ANALYZING SOCIAL RELATIONSHIP IN VISUAL SOCIAL MEDIA

by

Xiaolong Wang

A dissertation submitted to the Faculty of the University of Delaware in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Computer Science

Spring 2016

© 2016 Xiaolong Wang
All Rights Reserved
ANALYZING SOCIAL RELATIONSHIP IN VISUAL SOCIAL MEDIA

by

Xiaolong Wang

Approved: 

Kathleen F. McCoy, Ph.D.
Chair of the Department of Computer and Information Science

Approved:

Babatunde A. Ogunnaike, Ph.D.
Dean of the College of Engineering

Approved:

Ann L. Ardis, Ph.D.
Senior Vice Provost for Graduate and Professional Education
I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed:  
Chandra Kambhamettu, Ph.D.  
Professor in charge of dissertation

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed:  
Jingyi Yu, Ph.D.  
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed:  
Antwan D. Clark, Ph.D.  
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed:  
Noel C. Codella, Ph.D.  
Member of dissertation committee
ACKNOWLEDGEMENT

First, I would like to express my gratitude to my Ph.D advisor Professor Chandra Kambhamettu for his support and supervision during my study. During the past four years, he has taught me how to explore research topics and how to solve problems. He also gave me lots of encouragements. It is my honor to have Dr. Chandra Kambhamettu as my advisor, and I am lucky to be supported by him. I also want to thank Professor Jingyi Yu, Dr. Antwan D. Clark and Dr. Noel C. Codella for providing suggestions for my dissertation and serving as my committee members.

I would also like to thank all my previous and current labmates in VIMS including Abhishek Kolagunda, Guoyu Lu, Scott Sorensen, Philip Saponaro, Vicent Ly, Rohith MV, Stephen Rhein, and Wayne Treible for their help. I would also like to thank Dr. John Smith in IBM Watson research and Dr. Jon Currey in Samsung Research America for offering me internship opportunities. I am lucky to meet many good friends in University of Delaware including Benwen Zhang, Xiaoran Wang, Nan Zhou, Yu Ji, Ruoyao Ding, Gang Li, Jingwei Ye, Chen Huo, and Can Chen. I also thank my colleagues Rui Guo in The University of Tennessee, Knoxville, Yin Zhou in Apple and Michele Merler in IBM for their collaborations in my research projects.

Most of all, I am deeply grateful to my wife Lei (Erica) Wang. Her love, support and her optimism greatly inspired my passion for research. My lovely daughter Xixi (Daisy) Wang makes me enjoy the fun of being a father. Most importantly, I would like to dedicate this work to my parents, Zhongli Wang and Lixiang Feng. They encouraged me to seek my dream and be optimistic since my childhood. Without their love and support, I can’t make this possible.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>LIST OF TABLES</th>
<th>ix</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF FIGURES</td>
<td>xi</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>xv</td>
</tr>
</tbody>
</table>

## Chapter

1 **INTRODUCTION** .................................................. 1

1.1 Outline ......................................................... 4
1.2 Thesis Related Publications ............................... 5

2 **KINSHIP VERIFICATION** ........................................ 6

2.1 Introduction .................................................. 6
2.2 Related Work .................................................. 6
2.3 Proposed Methods ............................................... 8
  2.3.1 Appearance Feature Extraction ............................. 8
  2.3.2 Geometric Feature Extraction .............................. 11
  2.3.3 Classification ............................................... 13

2.4 Experiments ................................................... 13
  2.4.1 Database .................................................. 13
  2.4.2 Experimental Setup ......................................... 14
  2.4.3 Experimental Results and Analysis ........................ 14

2.5 Summary ....................................................... 15

3 **FAMILY PHOTO DETECTION** ...................................... 17

3.1 Motivation ..................................................... 17
3.2 Introduction .................................................. 17
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.3</td>
<td>Related Work</td>
<td>20</td>
</tr>
<tr>
<td>3.4</td>
<td>Our Approach</td>
<td>21</td>
</tr>
<tr>
<td>3.4.1</td>
<td>Geometry Model</td>
<td>22</td>
</tr>
<tr>
<td>3.4.2</td>
<td>Appearance Model</td>
<td>28</td>
</tr>
<tr>
<td>3.4.2.1</td>
<td>Pairwise Feature (PF) Extraction:</td>
<td>29</td>
</tr>
<tr>
<td>3.4.2.2</td>
<td>Facial Codebook Construction</td>
<td>33</td>
</tr>
<tr>
<td>3.4.2.3</td>
<td>DOGF Feature Extraction</td>
<td>34</td>
</tr>
<tr>
<td>3.4.3</td>
<td>Semantic Model</td>
<td>35</td>
</tr>
<tr>
<td>3.4.4</td>
<td>Classification</td>
<td>37</td>
</tr>
<tr>
<td>3.5</td>
<td>Experiments</td>
<td>38</td>
</tr>
<tr>
<td>3.5.1</td>
<td>Dataset</td>
<td>38</td>
</tr>
<tr>
<td>3.5.2</td>
<td>Experimental Results</td>
<td>40</td>
</tr>
<tr>
<td>3.5.3</td>
<td>Discussion</td>
<td>41</td>
</tr>
<tr>
<td>3.6</td>
<td>Conclusion</td>
<td>42</td>
</tr>
</tbody>
</table>

4 AGE ESTIMATION VIA SUPERVISED DEEP NEURAL NETWORK  

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>Background and Related Works</td>
<td>49</td>
</tr>
<tr>
<td>4.2</td>
<td>Related Work</td>
<td>51</td>
</tr>
<tr>
<td>4.3</td>
<td>Methods</td>
<td>54</td>
</tr>
<tr>
<td>4.3.1</td>
<td>Convolutional Neural Networks (CNNs)</td>
<td>55</td>
</tr>
<tr>
<td>4.3.1.1</td>
<td>Convolution Layers</td>
<td>56</td>
</tr>
<tr>
<td>4.3.1.2</td>
<td>Sub-sampling Layers</td>
<td>56</td>
</tr>
<tr>
<td>4.3.1.3</td>
<td>Training Process</td>
<td>57</td>
</tr>
<tr>
<td>4.3.2</td>
<td>Deep Learned Aging Pattern (DLA)</td>
<td>58</td>
</tr>
<tr>
<td>4.3.3</td>
<td>Regression or Classification</td>
<td>61</td>
</tr>
<tr>
<td>4.4</td>
<td>Experiments</td>
<td>61</td>
</tr>
<tr>
<td>4.4.1</td>
<td>Datasets and Experimental Settings</td>
<td>62</td>
</tr>
<tr>
<td>4.4.2</td>
<td>Experimental Results and Analysis</td>
<td>63</td>
</tr>
<tr>
<td>4.5</td>
<td>Summary</td>
<td>66</td>
</tr>
</tbody>
</table>
5 AGE ESTIMATION VIA CONVOLUTIONAL SPARSE CODING 70

5.1 Introduction 70
5.2 Proposed Approach 73

5.2.1 Convolutional Sparse Coding 74
5.2.2 Feature Extraction 76
5.2.3 Discriminative Feature Learning 78

5.3 Experiments 79

5.3.1 Evaluation Criteria 79
5.3.2 Datasets and Experimental Settings 80
5.3.3 Experimental Results and Analysis 82

5.4 Conclusion 87

6 HOW TO MINIMIZE INFLUENCE DUE TO GENDER AND ETHNICITY FOR AGE ESTIMATION 90

6.1 Introduction 90
6.2 Approach 93

6.2.1 Aging Feature Representation 94
6.2.2 Discriminative Aging Mapping 95
6.2.3 Correlation Learning 95

6.2.3.1 Canonical Correlation Analysis 96
6.2.3.2 Partial Least Squares 97

6.2.4 Classification(Aging Function) 99

6.3 Dataset 99
6.4 Experiment 100

6.4.1 Experiment Setting 101
6.4.2 Experimental Results Analysis 103

6.5 Conclusion 104

7 SUMMARY 106
## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Comparison of different methods on kinship verification.</td>
<td>15</td>
</tr>
<tr>
<td>3.1</td>
<td>Classification results using different schemes. Fusion Scheme I indicates the fusion scheme based on geometry and appearance information (CNN). Fusion Scheme II indicates the fusion model using appearance, geometry and semantic information together. Fusion Scheme III indicates the fusion scheme based on geometry and appearance information (dense SIFT based). Chen et al.(^*) indicates the assumption that approach [1] could classify the newly added images on the collected dataset with 100% classification accuracy.</td>
<td>39</td>
</tr>
<tr>
<td>4.1</td>
<td>Illustration of the used datasets. (N_{in}) indicates the number of individuals. (Age_{r}) represents the age range of the corresponding dataset.</td>
<td>61</td>
</tr>
<tr>
<td>4.2</td>
<td>Comparison of different methods for age estimation on two benchmarks (measured in MAE).</td>
<td>63</td>
</tr>
<tr>
<td>5.1</td>
<td>Illustration of the used datasets. (N_{in}) indicates the number of individuals. (Age_{r}) represents the age range of the corresponding dataset.</td>
<td>80</td>
</tr>
<tr>
<td>5.2</td>
<td>Comparison of different methods for age estimation on two benchmarks in MAE (years). Proposed scheme indicates STD pooling is applied in the proposed pipeline.</td>
<td>84</td>
</tr>
<tr>
<td>5.3</td>
<td>Comparison of MAE in Different Age Ranges on the FG-NET Database</td>
<td>87</td>
</tr>
<tr>
<td>6.1</td>
<td>The distribution of face images with different races and genders in Morph II database</td>
<td>100</td>
</tr>
<tr>
<td>6.2</td>
<td>Experimental Results: MAE(in yrs.). B+M+C represent BIF feature + MFA + CCA. B+M+P represents BIF+MFA+PLS. The age groups include African-American Female(BF), African-American Male(BM), Caucasian Female (WF). MAE Incr. indicates the percentage of advance using B+M+P compared to the result only based on BIF.</td>
<td>102</td>
</tr>
</tbody>
</table>
6.3 The distribution of images used in our experiment . . . . . . . . . . . . . 102
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Illustrations of facial images with kinship relation. The relation includes mother-daughter, mother-son, father-daughter, father-son and siblings.</td>
</tr>
<tr>
<td>2.2</td>
<td>The illustration of our proposed framework for kinship verification. (a) An illustration of appearance model. We construct the image pyramid based on the facial pair images. Then the feature extraction is conducted within overlapping patches. GMM is applied to find the corresponding similar patch pairs between two individuals. In the end, the absolute difference between two patch feature vectors is calculated as the kinship feature fed into the classifier. We use different colors to indicate different feature sets. We also give one corresponding similar face patch measured by our model. (b) An illustration of the geometry model. First, we extract facial landmarks from the facial image based on [2]. Then we project the landmarks to the Grassmann manifold. Geodesic distance is calculated as the feature to measure the difference of two shapes.</td>
</tr>
<tr>
<td>2.3</td>
<td>Kinship Verification Results. Proposed scheme indicates the fusion scheme based on appearance and geometry.</td>
</tr>
<tr>
<td>3.1</td>
<td>Examples of family photos (first row) and non-family photos (second row). The similarity of appearance, and the arrangement of people in a group photo reveal the type of the photo. In general, family members usually stand in a cascaded way and are more similar in their appearance.</td>
</tr>
<tr>
<td>3.2</td>
<td>An illustration of extracting geometry feature of the given group photo. In our geometry feature extraction pipeline, each vertex on the polygon corresponds to one face in the group photo.</td>
</tr>
<tr>
<td>3.3</td>
<td>Illustrations of constructed polygon based on individual arrangement in a photo based on our approach. From the constructed polygons, we can see that there is a significant difference between people arrangement of family photo and non-family photo in most cases.</td>
</tr>
<tr>
<td>3.4</td>
<td>The proposed architecture of CNN for learning face representation. To be concise, only three convolution layers are shown here.</td>
</tr>
</tbody>
</table>
3.5 An illustration of appearance feature extraction framework based on SIFT.

3.6 An illustration of the proposed DOGF feature extraction. To be concise, only one face pair is listed here. Other facial pairs all follow the same processing procedure. In the codebook generation step, two colors represent different classes (family and non-family), which are represented by green and blue. In each category, four codebooks are learned from corresponding pairwise feature set divided by different age gaps.

3.7 Top five most discriminative semantic concepts associated with family (top) and non-family (bottom) photos. Under each photo we report the score of the corresponding semantic model. Images with red border represent mislabelings of the classifier.

3.8 (a) Illustrations of correctly classified photos using the geometry model. (b) Examples of misclassified photos using geometry model.

3.9 (a) Illustrations of correctly classified group photos based on appearance model (DOGF). (b) Illustrations of the photos misclassified using geometry model and correctly classified by the fusion model based on geometry and appearance cues.

3.10 Illustrations of the photos correctly classified using the semantic model.

3.11 Performance comparison of different models. Fusion Scheme I indicates the fusion model using geometry and appearance information. Fusion scheme II indicates the fusion model using appearance, geometry and semantic information together.

3.12 (a) Performance comparison between mean distance and the proposed geometric feature scheme. Mean distance indicates that we calculate the center position of the polygon, then calculate the distance between the center to all the vertices in the polygon. Afterwards, the calculated mean distance is used as the feature to classify the photo category. (b) Performance comparison for different appearance models. No age indicates that there is no age information used in appearance model.

4.1 An illustration of the aging process of two different individuals. Each row are images of the same person at different ages.

4.2 Illustration of CNN structure for extracting deep learned age features.
<table>
<thead>
<tr>
<th>Section</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.3</td>
<td>Comparisons of different manifold learning algorithms in terms of MAE for age estimation with different dimensions in MORPH dataset.</td>
<td>64</td>
</tr>
<tr>
<td>4.4</td>
<td>Visualization of the first three components of the extracted deep learned feature after manifold learning (MFA) on the MORPH dataset. The age labels are colorized for a better view of the feature distribution after manifold mapping.</td>
<td>67</td>
</tr>
<tr>
<td>4.5</td>
<td>Cumulative scores of the algorithms with different settings on MORPH dataset.</td>
<td>68</td>
</tr>
<tr>
<td>4.6</td>
<td>The age estimation errors using different methods on MORPH dataset.</td>
<td>69</td>
</tr>
<tr>
<td>5.1</td>
<td>An illustration of the proposed framework. Before feeding into the proposed pipeline, images are first aligned based on eyes coordinates. Then a given image goes through the feature extraction pipeline consisting of convolution, Abs, LCN and STD pooling. Next, manifold learning is applied to the vectorized feature maps obtained from previous steps. Manifold learning is employed to transform the high-dimensional feature vector into a low-dimensional feature space, where dimension-reduced features are more discriminative for age estimation. Finally, regression model is used to predict the age label for the input image.</td>
<td>73</td>
</tr>
<tr>
<td>5.2</td>
<td>The learned $7 \times 7$ filters. We highlight a set of impressive filters that are semantically meaningful, e.g., nose (green), eye (red), lip (blue) and wrinkle (purple)-like filters.</td>
<td>75</td>
</tr>
<tr>
<td>5.3</td>
<td>Overview of age distribution of two datasets (Morph on the left and FG-NET on the right) used in our experiment. X-coordinate represents age label. Y-coordinates represents the number of people within each age category.</td>
<td>80</td>
</tr>
<tr>
<td>5.4</td>
<td>Example images from the FG-NET dataset. Numbers below the image indicate different ages.</td>
<td>81</td>
</tr>
<tr>
<td>5.5</td>
<td>Example images from the MORPH dataset. Numbers below the image indicate different ages.</td>
<td>82</td>
</tr>
</tbody>
</table>
5.6 (a) Illustration of the facial image distribution of one individual at different ages, on the age manifold for FG-NET dataset. (b) Illustration of facial image distribution of different individuals with different ages on the age manifold for the MORPH dataset. Different color represents different age categories. The number in red denotes the ground truth age of each facial image. 83

5.7 Correlations between age estimates by the proposed approach vs. ground truth ages on (a) Morph and (b) FG-NET. 85

5.8 Cumulative scores (at error levels from 0 to 20 years) of different schemes on MORPH. STD, AVE, MAX indicate standard deviation, average and max pooling strategy respectively. 88

5.9 Cumulative scores (at error levels from 0 to 20 years) of different schemes on FG-NET. STD, AVE, MAX indicate standard deviation, average and max pooling strategy respectively. 89

6.1 Example of facial images included in MORPH-II at age 30 with different genders and races. Face images in the first and second row are female, and last two rows are male. The images in the first and third row belong to the Caucasian (White) and the remaining images belong to African American (Black). 91

6.2 Illustration of the proposed learning scheme. Different colors represent the features extracted from different races (red for African American, and green for Caucasian). The race differences are marked by different shapes, which are circle and triangle. Same size corresponds to the same age label. Our scheme is separating different ages “further” away while keeping same age “closer” regardless of the race (or gender) difference. 92

6.3 The cumulative scores of age estimation based on different learning schemes. No cross means the training and testing sets are within the same race/gender age group. Except no cross, all other experiments are conducted between different age groups for training and testing. Each curve is the average of the experimental results of four ethnicity/gender age groups (WF, WM, BF, BM). 105
ABSTRACT

My thesis focuses on analyzing image-based social relations of people in a given photo. Compared to general face recognition problems, social relation analysis between facial images is a new research topic. There are two parts in my thesis: first part deals with pair-wise kinship verification, and second part deals with group photo analysis for social relation. Because of the importance of age estimation in group photo analysis, I propose a novel age estimation scheme to improve the estimation performance. Furthermore, I also study the scheme to minimize the influence of gender and race for age estimation.

Most current studies are based on measuring the appearance similarities between global faces. However, most significant cues for kinship verification are from local facial regions instead of the global face. My study solves this problem by finding these local appearance cues. Meanwhile, to improve the verification performance, I also build a 2D face geometry model to measure the face shape similarity between facial images. Comprehensive evaluations are performed for my fused appearance and geometry model.

In everyday life, people usually take photos together. Predicting the relationship of group of people thus becomes an essential topic for image-based social relation inference. It has several applications since family is one of the most important social units in the society, thus categorizing family photos constitutes an essential step towards image-based social analysis and content-based retrieval of consumer photos. Our model aims to capture the characteristics of group photos from different aspects. We propose an approach that combines multiple unique and complimentary cues for recognizing family photos. I first propose a geometry model to characterize people’s standing pattern. To better measure the facial
similarity, I apply deep neural network for measuring the appearance similarities between different individuals. I also apply the scene understanding knowledge to improve the group photo recognition performance. Experimental results demonstrate that age information plays an important role in the group photo categorization.

To improve the performance of age estimation, I propose a new aging feature extraction scheme via Convolutional Neural Network (CNN). The new aging pattern is learned through the image data instead of using manually-crafted features. My proposed aging pattern consists of the feature extracted from many hierarchical layers instead of only using the top-layer. To improve the age estimation performance and the efficiency, I apply manifold learning to project the extracted aging feature into another low-dimensional feature subspace. I also evaluate the performance using different regression and classification methods to predict the final age value.

One drawback of deep neural networks is its complicity and model size. This especially poses a challenge for deploying the learned model on mobile devices. To get a smaller model with a comparable performance, I advocate a new unsupervised aging feature via convolutional sparse coding. Experimental results demonstrate that our learning approach better extracts localized subtle aging features instead of losing them in a deep layered network like CNN, and also significantly reduces the memory consumption.

The performance of the age estimation system is usually influenced by many factors, such as expression, gender, and ethnicity. In this thesis, I present a new scheme to mitigate the influences due to race and gender in the problem of age estimation. I apply the correlation learning scheme on two different feature subspaces to learn the correlation projection matrix. Then I utilize discriminant learning on the projection subset learned by the correlation learning. As illustrated by experimental results, the proposed model improves the performance of age estimation across different age groups significantly.
Chapter 1

INTRODUCTION

Facial image analysis is one of the most important topics in computer vision and biometrics. Human facial images convey large quantities of visual information, such as identity, age, emotion, gender, ethnicity, etc. These types of information have very wide applications in many areas including security, law-enforcement, entertainment, human-computer interaction (HCI) system, and artificial intelligence system among others.

Besides obtaining these general attributes (e.g., gender, emotion, etc) from facial images, we can also extract other useful information, such as the kinship relation. Kinship is defined as a relationship between two people who have biological relation. Studies in psychology and biology have found that humans have the ability to identify kinship relations. I propose to verify the kinship relation via the measurement of familial traits obtained from facial images. Compared to face recognition, kinship verification is much more challenging. One reason is because the similarity between two facial images with kinship relation is very subtle. For example, the similarity between children and their parents is not always obvious. Even the resemblance between father-daughter and mother-daughter is different. Most recent work measures the similarity of two global faces to verify the kinship. However, direct global face matching is not very accurate in measuring the kinship verification since the similarities between two face images with kinship are usually on local face regions. Additionally, it is inferior in locating familial cues between two comparing facial images. I introduce a fusion framework with the integration of appearance and geometry verification schemes. My appearance model can help locate familial traits between two facial images.
My appearance model uses Gaussian Mixture Model (GMM) as the basis to locate and measure the similarity between familial traits. My proposed geometry model uses landmarks extracted from the facial images as the basis to measure the kinship relation between people. I also fuse the geometry and appearance models together to predict the kinship relation of two facial images.

As social media usage increases, a wide variety of information can be extracted from the growing number of consumer photos shared online, such as the category of events captured or the relationships between individuals in a given picture. Family is one of the most important units in the society, thus categorizing family photos constitutes an essential step toward image-based social analysis and content-based retrieval of consumer photos. I propose an approach that combines multiple unique and complimentary cues for recognizing family photos. The first cue analyzes the geometric arrangement of people in the photograph, which characterizes scene-level information with efficient yet discriminative capability. The second cue models facial appearance similarities to capture and quantify relevant kinship relations between two individuals in a given photo. The last cue investigates the semantics of the context in which the photo was taken. Experiments on a dataset containing thousands of family and non-family pictures collected from social media indicate that each individual model produces good recognition results. Furthermore, a combined approach incorporating appearance, geometric and semantic features significantly outperforms the state of the art in this domain. Experimental results also demonstrate that age information plays an important role in the family photo categorization, especially in the appearance model.

To improve the performance of age estimation, I propose a new aging feature based on convolutional neural network. Estimating age from human facial images is a traditional problem in computer vision and the biometric community. Inferring age attribute has many applications. These applications include security access control, mugshot identification, automatic photo labeling and entertainment. However, estimating age is very challenging since
aging information is difficult to capture and is influenced by many factors. To accurately estimate age from facial images, one key problem is to extract the effective aging feature. Instead of using traditional hand-crafted features, I propose to use supervised convolutional neural network to learn aging features from facial images. I tested the proposed approach in two benchmarks and obtained very promising results.

My CNN based aging feature has proved to be very powerful in capturing aging cues. However, deep CNN always has a large parameter space and memory requirements. This makes the deep convolutional architecture incompatible with mobile devices. How can we effectively leverage feature learning approach to improve the performance, but still avoid subtle feature loss and unnecessary memory budget? My solution is motivated by Convolutional Sparse Coding (CSC)’s recent success in computer vision and its unique advantages: CSC considers the whole image during training instead of the patch level, therefore the filter learned by CSC can capture high-order image statistics. Therefore, I employ CSC as the building block to automatically learn a semantically meaningful filter bank, in the hope that the learned filter bank can capture more comprehensive and accurate aging features. Furthermore, the aging estimation model based on CSC runs at a much cheaper memory and power budget.

The performance of age estimation is usually influenced by the variation of ethnicity and gender. Training and testing across different age groups can greatly decrease the performance of the general age estimation system. Compared to the large corpus on general age estimation problems, there are relatively few works specifically studying the influence of race and gender. In this work, I have studied the influence of age and gender in the age estimation topic and proposed a discriminant correlation-learning based solution for improving the performance. In this thesis, I mainly evaluate age estimation performance across two races - Caucasian and African American. I have conducted a set of comprehensive experiments to validate the proposed pipeline. From the experiment, my proposed pipeline can
mitigate the effect caused by gender and race variations. My scheme can reduce the estimated Mean Absolute Error (MAE) significantly. This also validates the feasibility of age estimation across different race/gender age groups. This study is a necessary addition for the current age estimation system.

Overall, I investigate several different problems. I mainly focus on analyzing social relation from facial images. For recognizing kinship relation, a fusion scheme based on appearance and geometry information is proposed. Not limited to individual facial images, I also present a new framework for recognizing family photos. The advocated framework characterizes the scene-level information (geometric arrangement of people), facial similarities and the semantics of the context of a given group photo. I also present two new aging feature extraction schemes. One is based on the supervised convolutinal neural network and anther one uses the convolutional sparse coding as the basis to learn aging feature. In addition, I also study the influence of gender and ethnicity to the age estimation system and propose a discriminant correlation learning based solution.

1.1 Outline

In Chapter 2, I discuss the kinship verification problem and investigate the approach for locating familial traits from facial images and utilizing geometry cues help improve the verification performance. In Chapter 3, I present a framework for family photo detection and discuss the geometry and appearance model in details. Chapter 4 focuses on the supervised age estimation approach. In this chapter, the related work for age estimation is also discussed. I evaluate the proposed approach on two popular benchmarks. In chapter 5, I study the unsupervised scheme for extracting aging features. In Chapter 6, I analyze influential factors for age estimation due to gender and ethnicity and propose a new scheme to deal with these influences. In the end, I summarize the thesis.
1.2 Thesis Related Publications


Chapter 2

KINSHIP VERIFICATION

2.1 Introduction

One goal of computer vision is to enable a computer to automatically extract semantic information and perceive knowledge from an image or image sequences. Although various techniques have been developed in the past few years, it is still a great challenge, especially when compared to human level proficiency. An automated family relationship analysis system is essential for annotation of social relations in photos. This technique is also useful for intelligent social computing or social media understanding.

Compared to face recognition technology, the kinship verification problem is much more challenging. One reason is the subtle similarity between two facial images with a kinship relation. For example, it is not easy to capture the similarity between children and their parents. Considering the addition of variations of pose, expression, etc, the kinship verification problem becomes much more complicated.

2.2 Related Work

Humans are capable of recognizing family members. Human perception of kinship has been studied in psychology and evolutionary studies as an active area of research [3, 4, 5]. In the computer vision area, this problem is first studied in [6], where color and distance between facial parts are used as features to determine the kinship relation. The features that they used include eye color, distance between eyes and nose, etc. Then k-nearest neighbor (KNN) is applied as the classifier. In [7], Xia et al. applied transfer subspace learning approach to minimize the difference between children and old parents via young
Figure 2.1: Illustrations of facial images with kinship relation. The relation includes mother-daughter, mother-son, father-daughter, father-son and siblings.

parents as the intermediate class. They also used Gabor filters to extract features for kinship verification in an extended dataset in [8]. Naman et al. [9] applied self-similarity descriptor (SSD) to represent the similarity between facial images to verify the kinship relation. The ensemble metric learning scheme is utilized in [10]. Guo and Wang employed the matching score obtained from facial parts (eyes, mouth and nose) to determine the relation between two facial images in [11]. In work [12], a new metric learning scheme was applied to move samples within intra class (kinship relation) closer and separate the samples of inter class (without kinship relation) further. Instead of using images, expression videos in recent work [13] were used to verify kinship based on resemblance of facial expressions, where facial dynamics and spatio-temporal appearance features were fused together to determine the relation.

Our approach mainly includes the following novelties: 1) We propose an effective feature matching scheme to locate familial traits of two facial images. This scheme demonstrates significant improvement over previous works. 2) Our approach also utilizes shape information for kinship verification. Experiments illustrate that facial geometric feature is also discriminatory for determining kinship relation. 3) We fuse appearance and geometry
information together. The advocated fusion scheme has a significant improvement over the baseline by more than 17%. We organize the structure of this chapter as follows: the proposed scheme is illustrated in Section 2.3. Experimental results and analysis are presented in Section 2.4. In Section 2.5, we summarize the whole chapter.

2.3 Proposed Methods

Our goal is to explore an efficient automatic scheme to detect kinship cues and verify the kinship relation between two faces using the aforementioned cues. In this section, we give the details of the proposed scheme. The general scheme is illustrated in Figure 2.2 where appearance and geometry models are included. Before the analysis of the fusion model, we demonstrate appearance and geometry model separately. The fusion scheme is based on the results obtained from appearance and geometry models.

2.3.1 Appearance Feature Extraction

In general, most works for kinship verification use global face [12] or facial parts [11, 14]. However, because of imperfections in face alignment and the influence of pose, expression, occlusion, etc., there are lots of variations in the aligned face. In our framework, we design a robust feature matching scheme to locate kinship cues to calculate the similarities between two facial images. The proposed scheme can minimize the effects of pose and expression variations, improving verification performance.

We first build a three-layer Gaussian image pyramid for each image as illustrated in Figure 2.2. We then apply Local Binary Pattern (LBP) [15] to extract the feature vector within overlapping patches. Other features can also be used, such as SIFT [16], HOG [17], etc. Let us assume that there are $N$ patches extracted from a face image, the feature set corresponding to this image can be represented as $F = \{f_i\}_{i=1}^{N}$, where $f_i$ indicates the feature extracted from the $i$th patch, $N$ is the total number of patches.
Figure 2.2: The illustration of our proposed framework for kinship verification. (a) An illustration of appearance model. We construct the image pyramid based on the facial pair images. Then the feature extraction is conducted within overlapping patches. GMM is applied to find the corresponding similar patch pairs between two individuals. In the end, the absolute difference between two patch feature vectors is calculated as the kinship feature fed into the classifier. We use different colors to indicate different feature sets. We also give one corresponding similar face patch measured by our model. (b) An illustration of the geometry model. First, we extract facial landmarks from the facial image based on [2]. Then we project the landmarks to the Grassmann manifold. Geodesic distance is calculated as the feature to measure the difference of two shapes.
A Gaussian mixture model is composed of a weighted sum of $T$ component Gaussian densities. It can be computed as:

$$P(v|\lambda) = \sum_{t=1}^{T} w_t g(v|u_t, \Sigma_t),$$  \hfill (2.1)

where $v$ is $D$ dimensional feature vector, $w_t$ indicates the weight associated with each Gaussian density, and $\sum_{t=1}^{T} w_t = 1$. $g(v|u_t, \Sigma_t)$ is the component Gaussian density function with mean vector $u_t$ and covariance matrix $\Sigma_t$. To compute the parameters of GMM, we use the Expectation Maximize (EM) algorithm. It contains two iterative steps, which are E-step and M-step. E-step calculates the log-likelihood of the distribution, and M-step is used to update the parameters to maximize the log-likelihood [18].

GMM has many applications, such as scene recognition [19], speech recognition [20], video face recognition [21] and biometrics fusion [22]. In this work, we use GMM to locate the kinship cues between two given faces. After obtaining the GMM model based on the training set, we use GMM scheme to locate the most similar appearance regions corresponding to facial image pairs, then extract the similarity feature difference for classification. Because each component of GMM denotes one particular appearance characteristics in one facial area, it can be used to locate the facial area sharing the most similar cues. Each facial image corresponds to one feature set $F = \{f_i\}_{i=1}^{N}$. After we get GMM with $T$ components based on the training set, feature vector $f_i$ obtained in each patch is plugged into these $T$ Gaussian components with the associated weight $w_t$. Then we can find that the corresponding $k$th patch which gets the largest probability value based on

$$\max_t w_t g(v|u_t, \Sigma_t).$$ \hfill (2.2)

Given $T$ components of the GMM, we obtain $T$ feature vectors to maximize the value of the
corresponding GMM component based on Eqn. 2.2. We name this feature vector GMM as patch feature. This can be represented as \( F_g = \{ f_{g1}, f_{g2}, \ldots, f_{gT} \} \). Given facial images \( a \) and \( b \), their corresponding GMM patch features are \( F_{ga} \) and \( F_{gb} \). We calculate the absolute difference of \( F_{ga} \) and \( F_{gb} \) to represent the similarity of two facial images. In some sense, the similarity measurement between two facial images can be represented as \( S = |F_{ga} - F_{gb}| \).

Since \( F_{ga} \) and \( F_{gb} \) are composed of patch features which maximize the probability of the same Gaussian component, \( S \) can fully represent the likeness between the corresponding features of two facial images. In our experiments, facial images are unconstrained, such as illustrated in Figure 6.2. There are many variations, such as pose, expression, etc. Our experimental results demonstrate the effectiveness of the proposed scheme for feature matching.

### 2.3.2 Geometric Feature Extraction

In the proposed scheme, we also use geometric information extracted from facial image to verify the kinship relation. As discussed in [4], family members share similar facial shape and appearance. Such as illustrated in Figure 2.2, the shape of the chin and mouth of the daughter and mother are very similar.

We measure the similarity of 2-D geometry of the facial shape to verify the kinship relation between two individuals. In the first step, 68 facial landmarks are extracted from the facial image. We use the landmark detection algorithm proposed in [2]. Next, we apply Grassmann manifold to represent the kinship similarity given these facial shapes. Grassmann manifold is widely used in image processing and computer vision [23, 24, 25, 26]. Facial landmarks have been widely used to represent the facial shape. We denote the landmark coordinates associated with a given face as \( X = [(x_1, y_1), (x_2, y_2), (x_1, y_1), \ldots, (x_p, y_p)] \).

As we know, the obvious drawbacks of directly using the facial landmarks for matching are the sensitivity to changes of pose, view, etc. To minimize the influence of these factors, we extract the tall-thin orthonormal matrix to represent the given subspace. We apply SVD to \( X \),
that is \( X = A \Sigma B \). Any arbitrary face shape can be considered as the spatial transformation of the base face. It means if two facial shapes correspond to the same basis shape, they will span the same subspace. \( Z = AA^T \) is used as the Grassmann representation of \( X \).

A Grassmann manifold \( G_{m,n} \) represents the space where points are \( n \)-dimensional linear subspace in \( \mathbb{R}^m \). The facial shape can be constructed using \( p \times 2 \) matrices. To apply Grassmann manifold to represent the facial shape, we have \( m = p, n = 2 \). In this work, geodesic distance is used to measure the difference between two sets of points lying on the Grassmann manifold [27]. Grassmann manifold can be modeled as the quotient group of orthogonal groups \( SO(n) \) [28]. This indicates that geodesics in \( G_{m,n} \) can be represented using one-parameter exponential flow \( t \rightarrow exp(tE) \), where \( E \) is a skew-symmetric matrix. It is constructed as:

\[
E = \begin{pmatrix}
0 & D^T \\
-D & 0
\end{pmatrix},
\]

(2.3)

where sub-matrix \( D \) determines the direction and speed of the geodesic flow. Given a point \( X_0 \) on the Grassmann manifold indicated by its orthonormal basis \( Z_0 \), the geodesic path starting from \( S_0 \) is specified by \( \Phi(t) = Q \exp(tE) J \), where \( Q \in SO(n) \), \( Q^T Z_0 = J \) and \( J = \left[ I_k, 0_{m-k,k} \right] \). Projection matrix \( Z_0 \) is the representation of \( X_0 \) on the Grassmann manifold. The specific details can be found in [29]. Extrinsic description is employed to measure the similarity between two facial shapes. We can calculate the idempotent projection matrix as \( Z = AA^T \). Given two facial shapes \( X_1 \) and \( X_2 \), their corresponding projection matrix are \( Z_1 \) and \( Z_2 \).

Based on [29], the geodesic distance between them can be calculated as:

\[
Z_1 = \exp(tY) \ Z_2 \ \exp(-tY).
\]

(2.4)
The matrix $Y$ can be computed via eigen-decomposition of matrix $\chi = Z_1 - Z_2$ [29]. In our framework, $\chi$ is used as the feature vector to represent the similarity between two facial shapes, since $\chi$ computes the geodesic between the comparing two shape.

### 2.3.3 Classification

Kinship verification can be categorized into a binary classification problem. Given a pair of facial images, we aim to determine if the pair has a kinship relation or not. In this framework, we use Support Vector Machine with the Gaussian Radial Basis function (RBF) for its good performance in binary classification problem. We calculate similarity features from the appearance and geometry models and then fed each into individual SVM classifier. Our fusion scheme is based on the results obtained from appearance and geometry classifiers. In fusion model, an additional SVM classifier is used. The classifier of fusion scheme is trained using the intermediate scores obtained from the two schemes. We employ LibSVM [30] to train each SVM classifier. To get the optimized parameters, we follow five-fold cross-validation technique as discussed in [30].

### 2.4 Experiments

In this section, the experimental results of the proposed scheme are discussed. We also compare the proposed scheme with previous works in kinship verification. The advocated scheme is evaluated on a recently collected kinship database.

#### 2.4.1 Database

In this work, we use a recently collected unconstrained database “Family 101”[31] to evaluate the performance of the proposed algorithm. This dataset has 607 individuals which are collected from 206 different families. It is one of the largest current databases used for kinship verification. To label the relations of the whole dataset, Amazon Mechanical Turk was applied. The images are all collected from the internet in an uncontrolled environment.
The collected images have expression variations and pose changes. Occlusion is also included. The general illustration of the whole dataset is illustrated in Figure 6.2. Moreover, this database also has the diversity of different races, genders and ages. In our experiment, we randomly select 320 different facial pairs with kinship relation, and 320 facial pairs without kinship relation to construct the whole dataset.

2.4.2 Experimental Setup

In our experiment, we employ five-fold cross-validation. There is no overlaps between the training and testing. It means if an image is used in training, it is not used for testing on that fold. All images are cropped to the same size (120 × 150) before feature extraction. We set the scaling factor of Gaussian image pyramid to 0.5. LBP feature extraction is conducted in patches of 16 × 16 with 8-pixel spacing in each scale of image pyramid.

In appearance model, GMM is applied to extract the matching patch features from image pair. Then we calculated difference vectors. SVM with RBF kernel is used for classification. In geometry model, Geodesic difference is calculated as the feature to represent the shape similarities. First, we calculate the performance using these two different models. Then we tested the fusion scheme integrating these two models. In addition to evaluate the performance of the proposed scheme in kinship verification, we also compare the proposed framework with other approaches [6, 11, 10]. We list the comparison result in Table 2.1 and Figure 2.3.

2.4.3 Experimental Results and Analysis

To give a better visualization of the performance of different matching schemes, we draw ROC curves. We illustrated this in Figure 2.3. The performance of different schemes are illustrated in Table 2.1. Our approach achieves 92.03% accuracy. Overall, the proposed method presents a significant improvement compared to the state-of-the-art by more than 17%. The images used in this experiment are not high-resolution. This has restricted the
Table 2.1: Comparison of different methods on kinship verification.

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fang et al. [6]</td>
<td>74.06%</td>
</tr>
<tr>
<td>Guo &amp; Wang [11]</td>
<td>73.44%</td>
</tr>
<tr>
<td>Somanath &amp; Kambhamettu [10]</td>
<td>72.50%</td>
</tr>
<tr>
<td>Geometry model</td>
<td>65.16%</td>
</tr>
<tr>
<td>Appearance model</td>
<td>86.09%</td>
</tr>
<tr>
<td>Proposed scheme (fusion)</td>
<td>92.03%</td>
</tr>
</tbody>
</table>

performance of [11]. Modified Hausdorff distance was computed as the matching scheme between facial parts in this work. Our results demonstrate the advantage of GMM over modified Hausdorff distance scheme at locating the familial traits in low resolution images. Our approach also obtains more satisfactory verification performance than [6, 10]. Simply using the appearance alone outperforms all baselines by a large margin. With the fusion of the geometry model, an additional 5% improvement is obtained. This result demonstrates that shape information in the human face still conveys useful information for kinship relation.

2.5 Summary

In this chapter, we have demonstrated that the familial traits can be learned from pairs of facial patches using GMM. The learned familial cues perform well at forming a decision for kinship verification. Compared to existing works, the proposed scheme improves the performance significantly. Furthermore, we have studied the facial geometry for kinship verification based on the Grassmann manifold. Integrating the geometry with the appearance information gives us more than 17% improvement over existing techniques.
Figure 2.3: Kinship Verification Results. Proposed scheme indicates the fusion scheme based on appearance and geometry.
Chapter 3

FAMILY PHOTO DETECTION

3.1 Motivation

As described in the previous chapter, analyzing social relations through image processing is an active and emerging research topic. However, most current research work focuses on analyzing the pair-wise relationship between two facial images. To analyze the possible relations in the photo with three or more people is still an open problem.

3.2 Introduction

Along with the popularity of mobile cameras and access to the internet everywhere in the world, people constantly capture millions of photographs and share through the social media. As such, there is a great demand for automatic labeling of photos to navigate and perform retrieval on large scale photograph collections. One of the most important applications is personal photograph classification. It covers a broad spectrum of concepts and entities, from people to activities, from scenes to food or abstract concepts. In this chapter, we focus on analyzing the social relationships among the subjects of a photograph.

A broad classification of photographs in personal albums would be that of family photos and non-family photos. A quick search on Flickr with the keyword “family” confirms that people not only possess family photographs but are keen on labeling them as such. This has inspired us to create an automatic method for classifying images (containing groups of people) into family and non-family photographs.

Moreover, recognizing and categorizing family photos is an essential step towards image-based social analysis. Our work can also facilitate the research work in the social
sciences. Since family analysis is one of the key problems in social science fields. An automatic recognizing family photos scheme can help enlarge their dataset more efficiently.

As illustrated in Figure 3.1, there are several typical differences between family and non-family photos. First, people with kinship relations usually share similar facial attributes and the degree of similarity is usually much more evident than people in non-family photos. Second, the spatial distribution of physical standing pattern of individuals in two different photo categories is distinctive. For instance, people in family photo usually exhibit irregular height distribution and stand in a cascaded pattern. Meanwhile, the age gap between people in family photos is usually bigger than people in non-family photos. There are also some unique phenomena of people’s standing patterns in family photos. For example, children usually stand in the front of their parents or relatives. Elders tend to stand in the center. For comparison, friends in non-family photos stand in random patterns and the age variation lies in a smaller range. Part of the reason is because people tend to make friends within their age group. Third, the context information also provides additional cues for the group photo categorization. For example, most family photos are not likely to be taken in a bar or club, whereas it is common for friends to snap pictures together in such environments. Based on these observations, we utilize multiple cues to build a comprehensive model to recognize family photos. In general, our model fuses different kinds of discriminant features in a unified framework.

The novelties of our work include the followings: (1) We propose a new geometry feature to capture people’s standing pattern at the scene level. Our proposed geometry feature extraction scheme purely uses the relative position of people in the image. The geometry model does not use age, gender or appearance information. This is different from previous works [1, 32]. Our geometry feature represents the relative position of individuals in the photo. The geometry representation is very efficient. Our experimental results demonstrate that the geometry model can obtain more than 87.0% classification accuracy on
Figure 3.1: Examples of family photos (first row) and non-family photos (second row). The similarity of appearance, and the arrangement of people in a group photo reveal the type of the photo. In general, family members usually stand in a cascaded way and are more similar in their appearance.

its own. (2) To better characterize facial similarities between different individuals in a group photo, we study two different schemes to represent appearance features. The first one is based on dense SIFT and the second one uses deep convolutional neural network. The deep model is trained on the FaceScrub dataset [33]. Our experimental results demonstrate that deep neural network yields improved recognition performance in comparison to using hand-crafted features. (3) We also use semantic information to differentiate two different photo categories. Finally, we propose a fusion scheme which combines all three schemes together. This helps yield even further improvements to performance. To evaluate the performance, we conducted experiments on a dataset containing thousands of group photos from Flickr. The three different representations prove efficient for the group photo classification task and exhibits complimentary information. The experimental results demonstrate that fusing geometry, appearance model, and semantic context information yields an improvement of 3.3% over the current state-of-the-art for family photo classification.
3.3 Related Work

Many meaningful social context information can be extracted from the group photos as studied in [34]. Gallagher and Chen also found that contextual features can help interpret demographic information, such as people’s age and gender. Their work also demonstrated that using the context information can help improve the performance of event recognition. Meanwhile, this work showed that context information acquired from the group photo can also aid demographic information perception. Not only estimating general demographic information (age and gender), pairwise relation between individuals also can be extracted as studied in Singla et al. [35]. In this work, the authors apply rule-based Markov Logic Network to detect and identify the social relation of different individuals. They formulated different constraints in the proposed framework, such as parents are older than the children. They used MC-SAT algorithm to combine hard and soft constraints to predict relationship between different persons. Wang et al. [32] compute pairwise facial features to identify person and estimate social relations. Their proposed pairwise features are extracted from each facial pair. These social context features include pairwise height difference, age gaps and closeness. Their results demonstrated that social feature helps improve the recognition performance. Recently, Chen et al. [1] built a sub-graph learning based approach for group photo classification. Their works tried to divide the group photos into two general categories: family photo or non-family photo. In their scheme, they used subgraphs to represent social relationships. They utilized age, gender, and pairwise distance construct the social subgraphs’ set. The feature of a given group photo is represented by calculating the distribution of extracted subgraphs corresponding to the pre-trained subgraph set. For classification, SVMs [36] was used as the classifier to determine the category of the given photo. However, their model did not use facial similarities between people. Experimental results demonstrate that when people have similar age and gender, their proposed framework may not work well, such as the classification between photos with siblings.
In this chapter, we target the family photo recognition problem. We aim to classify the given photo into family and non-family categories. We advocate a new geometry feature which can capture the global standing pattern of individuals in a photo. This is different from measuring individuals’ pairwise distance. Our geometry model is also very efficient and simple to implement since only location information is used in our model. Although its simplicity, the model still obtains very good performance. In addition, we also advocate a new appearance feature to represent the facial similarities of the group photo. We use convolutional neural network to train our appearance model. The new appearance feature has outperformed SIFT feature [16] based appearance model. Only using appearance model (deep neural network based) can obtain more than 90.0% classification accuracy. Furthermore, not limited to low-level and middle-level photo representation, we also use high level semantic features to discriminate different image categories. As far as we know, this is the first time that semantic feature is used in the family photo analysis task. To fully represent image features, we fuse these different models together to improve the recognition performance. Experimental results show that fusing multiple cues extracted from group photo can help improve the recognition performance.

The structure of the whole paper is organized as follows: the proposed framework is presented in Section 3.4, where three different models are introduced. Section 3.4.1 presents the geometry model. The appearance model is described in Section 3.4.2. Section 3.4.3 gives the details of the general framework for extracting contextual semantic information. The experimental setting and results are presented in Section 3.5, and finally we give conclusion remarks.

3.4 Our Approach

In this work, three different cues are extracted from a group photo to differentiate two different family categories. We also apply a fusion scheme to fuse these cues together.
These cues aim to characterize the photo in different aspects. Fusion results demonstrate that these features are complementary to each other. These models also help us have a further understanding of the group photo classification problem.

3.4.1 Geometry Model

In general, specific spatial arrangements of individuals in the photo are correlated with their relationship. These spatial patterns exhibited in the photo can be represented by many different ways, such as by measuring the physical distance [1], or computing variations in height [32], even the physical proximities in the geometry of the people standing in a photo (such as pairwise distance between people). For example, as illustrated in Figure 3.1, the height differences and the relative position of individuals vary between two different categories. We can infer the human relationship information from these height difference, for instance, there is usually a distinct height difference between the parents and their children.

As indicated in Figure 3.1, the height differences and the relative position vary between family and non-family photos. These height differences can indicate sort kind of human relationship information, such as there is usually a distinct height difference between the parents and their children. The standing position in the photo can also reveal their social roles: for example, grandparents in the family photo usually stand in the group center, whereas the children usually stand in front of the parents in most situations. Contrary to family photo, the height difference of people in non-family photo is usually smaller than family photo. The physical proximity of different persons within the non-family photo is often different from family photos. Using the pairwise distance and height difference between people as the social context feature is generally used in most works. For example, Wang et al. [32] computed the Euclidean distance between facial pair to construct social features. Gallagher and Chen [34] calculated the distance between faces’ centroid and each particular face as one of the contextual features to estimate the demographic information. Chen et
Figure 3.2: An illustration of extracting geometry feature of the given group photo. In our geometry feature extraction pipeline, each vertex on the polygon corresponds to one face in the group photo.

al. [1] calculated the number of people between two comparing individuals as the pairwise distance measurement to build the social subgraphs. Different from these approaches, our geometry feature is extracted at the global scene level to capture the whole standing pattern of the group people. Our proposed geometry feature can be directly applied to the photo categorization problem. The process of extracting proposed geometry feature extraction is as follows:

**Polygon Construction:** We use people’s location in the photo as the basis to build our polygon. Our geometry feature is calculated using the constructed polygon. The whole geometry feature extraction pipeline is shown in Figure 3.2. We use face locations of individuals in the group photo as the vertices to formulate the polygon. First, we apply the convex hull [37] algorithm to all vertices to construct the primary polygon. As we know, convex hulls have been used in many areas, such as in computer visualization [38], path planning [39], shape matching [40], crystallography [41], cartography [42], etc. In our work, convex
Figure 3.3: Illustrations of constructed polygon based on individual arrangement in a photo based on our approach. From the constructed polygons, we can see that there is a significant difference between people arrangement of family photo and non-family photo in most cases.

The convex hull is set as the basis to extract the geometry feature. The general idea of convex hull is that given a set of points $S$ in the Euclidean space, the convex hull is the smallest convex set that includes $S$. For each point $s_i$ in $S$, it assigns a non-negative weight $w_i$. All non-negative weights sum to one. This can be formulated as follows:

$$\left\{ \sum_{i=1}^{n} w_i s_i | (\forall i : w_i \geq 0) \wedge \sum_{i=1}^{n} w_i = 1 \right\}, \quad (3.1)$$

where $n$ is the number of points in set $S$. Since photos used in our work are collected from unconstrained environment, there is a high degree of variations of people in the group photo. Simply building the constructed polygon using convex hull cannot guarantee the vertices of the constructed polygon include all faces. The polygon built by the convex hull using points
set $S$ has the minimal perimeter. In general, some points of set $S$ are lying on or within the constructed polygon. Let us assume that one photo includes $n$ different individuals. After applying the convex hull, we can build a polygon with $m$ sequential vertices enclosing these $n$ points, where $m \leq n$. Because we want to characterize the global geometry shape of the standing pattern of all $n$ people in the photo, our next step is to add these remaining $n - m$ vertices to the constructed polygon.

Let us assume that the coordinates of sequential vertices of the polygon constructed by the convex hull are $S = [(u_1, v_1), (u_2, v_2), \ldots, (u_m, v_m)]$. The $n - m$ remaining vertices can be represented as $S' = [(u'_1, v'_1), (u'_2, v'_2), \ldots, (u'_{n-m}, v'_{n-m})]$. To include these $n - m$ remaining vertices to the current polygon without abruptly changing the structure of the constructed polygon, for each point in $S'$, we compute its two closest sequential vertices in set $S$. Then the new polygon is built based on their relative positions. Each point $(u'_i, v'_i)$ can be considered as the node lying between two adjacent points $(u_j, v_j)$ and $(u_{j+1}, v_{j+1})$ in set $S$. This indicates that any point $(u_g, v_g)$ in the 2D space can be represented by

\[
\begin{align*}
  u_g &= \alpha \cdot u_j + (1 - \alpha) \cdot u_{j+1}, \\
  v_g &= \alpha \cdot v_j + (1 - \alpha) \cdot v_{j+1}, 
\end{align*}
\]  

(3.2)

where $\alpha \in [0, 1]$. To locate the nearest vertices in set $S$ of the given point $(u'_i, v'_i)$, we use the Euclidean distance as the measuring criteria. We calculate the distance between $(u'_i, v'_i)$ in $S'$ and $(u_g, v_g)$ by

\[
D = \sqrt{(u'_i - u_g)^2 + (v'_i - v_g)^2}. 
\]  

(3.3)

Then we compute the derivative of $D$ with respect to $\alpha$ to find its two nearest neighbors lying on the constructed polygon for the given point $(u'_i, v'_i)$. The derivative of $D$ with respect to $\alpha$ can be calculated by $\frac{\partial D}{\partial \alpha} = 0$. Based on this calculation, we can obtain $\alpha$. $\alpha$ can be
represented by

\[ \alpha = \frac{(u_j - u_{j+1})(u_g - u_{j+1}) + (v_j - v_{j+1})(v_g - v_{j+1})}{(u_j - u_{j+1})^2 + (v_j - v_{j+1})^2}. \] (3.4)

After obtaining \( \alpha \), we can compute the distance \( D \) between the \((u_i', v_i')\) and the sequential vertices in \( S \) based on Eqn. 3.3. Though all the calculated distance \( D \), we can get the minimum distance \( D_{\text{min}} \). The minimum distance \( D_{\text{min}} \) can help us locate these two sequential corresponding vertices in \( S \). We name the this computing step as “convex hull post-processing”. We use this post-processing step to build the comprehensive polygon where all faces are located as vertices on the polygon. To better illustrate the result of the whole procedure, we have listed several examples in Figure 3.3. From these examples, we also can find that people’s standing difference between two categories truly exists.

**Geometry Feature Extraction**: After building the polygon, we can use it as the basis to extract geometry feature. In this work, the extracted geometry feature is directly used to characterize people’s standing pattern. Our proposed mid-level geometry feature is based on Fast Fourier Transform (FFT) [43]. The whole feature extraction framework proceeds as follows:

(a) Compute the center of the 2D polygon.

\[
u_{\text{cen}} = \frac{1}{n} \sum_{i=1}^{n} u_i,
\]

\[
v_{\text{cen}} = \frac{1}{n} \sum_{i=1}^{n} v_i,
\] (3.5)

where \((u_{\text{cen}}, v_{\text{cen}})\) demonstrates the centroid of the polygon. \((u_i, v_i)\) is the coordinate of vertices on the polygon. \(n\) is the number of vertices (number of individuals).
(b) After obtaining the centroid of the polygon, we divide the angle around the polygon center into $K$ folds evenly. This process can be represented as

$$\theta_k = \frac{2\pi}{K} \times k, \; k \in \{1, 2, ..., K\}. \quad (3.6)$$

angle $\theta_k$ corresponds to the $k$th ray going through the center to the edge of the polygon. $K$ is the total number of rays casting from the centroid. In this work, $K$ is set to 64.

(c) Given the origin $(u_{cen}, v_{cen})$ with angle $\theta_k$, we cast $K$ rays from the origin with an equal angular interval. Then the distance between the perimeter of the polygon and the center is calculated along with the specified direction $\theta_k$. As shown in Figure 3.2, $L(\theta_k)$ is the distance between the original centroid $(u_{cen}, v_{cen})$ and the perimeter along the ray oriented along $\theta_k$.

(d) After obtaining the ray vector $L$, we calculate Fast Fourier Transform (FFT) for each ray. Then we calculate the amplitude spectrum and normalize the value. We use the amplitude as the feature. This can be invariant to many influences, such as image rotation and the shift in the order of polygon vertices in the image. Normalization can also deal with variations in photos’ resolution. We select top $T$ amplitude spectrum values and use them the geometry feature to represent the geometry information for a given group photo. From our experimental results, $T$ can be set to a value around 50 typically. Our geometry feature can characterize the different spatial distribution of people across photos with the variation in resolution and orientation.

As analyzed above, there is a distinct difference between our geometry model and previous approach used geometry model in object recognition [44] and face recognition [45, 46] where the shape information is mainly used to characterize the geometry of one single object. Whereas, our geometry feature is used to encapsulate the global view of the people in the group photo.
Figure 3.4: The proposed architecture of CNN for learning face representation. To be concise, only three convolution layers are shown here.

Figure 3.5: An illustration of appearance feature extraction framework based on SIFT.

3.4.2 Appearance Model

One important factor for measuring the facial similarity of people in the group photo is the appearance feature. In this work, I have studied two different kinds of features to characterize the appearance similarity. The first appearance feature measurement scheme is based on one of most successful hand-crafted feature SIFT. The second scheme is using deep neural network to extract appearance feature. In the beginning, we conduct face detection technique [47] to detect faces in the given photo and apply fiducial point detection algorithm [48] to the given face. Then we align all face images based on eyes’ coordinates. In
general, there are three different steps in the whole appearance model. They are Pairwise Feature (PF) Extraction, codebook construction and Degree Of Group similarity Feature (DOGF) feature extraction.

3.4.2.1 Pairwise Feature (PF) Extraction:

**SIFT feature based scheme:** Based on fiducial points, we divide the whole face into four facial parts: left eye, right eye, nose and mouth. Feature extraction is then conducted within each part. SIFT [49] is chosen as our feature for its good characteristics - robustness to the scale and affine transformation. Scale invariance is especially helpful when matching between young children and parents.

The SIFT feature is extracted with dense sampling in each facial part, which is illustrated in Figure 3.5. Then, to represent the facial similarity between two feature sets is our primary task. Let $A = \{ \vec{a}_1, \cdots, \vec{a}_n \}$ and $B = \{ \vec{b}_1, \cdots, \vec{b}_m \}$ denote two feature sets of the same facial part from two persons, $n$ and $m$ are the number of interesting points in sets $A$ and $B$, respectively. For faces captured in an unconstrained environment, the occlusion, pose variation, expression, etc. always affect the matching performance. To reduce the influences due to these factors, one may enlarge the facial parts, then use pyramid histogram matches or spatial pyramid matches [50]. In this work, we use a matching scheme adopted from the Modified Hausdorff distance (MHD) [51], which is typically used to measure the similarity between edge pixels. The Hausdorff distance is defined as $H(A, B) = \max(h(A, B), h(B, A))$, where

$$h(A, B) = \max_{\vec{a} \in A} \min_{\vec{b} \in B} ||\vec{a} - \vec{b}||$$ (3.7)
The MHD is calculated by:

$$h(A, B) = \frac{1}{A} \sum_{\vec{a} \in A} \left( \min_{\vec{b} \in B} \| \vec{a} - \vec{b} \| \right)$$  \hspace{1cm} (3.8)

In our work, instead of dealing with edge pixels [52], the MHD is used to measure the similarity between appearance features. This can also diminish the effects due to occlusion and pose variations in the face. First, we try to find the closest feature point $\vec{b}_0$ in $B$ corresponding to all the feature points $\vec{a}$ in $A$ based on the calculation of Euclidean distance. $\vec{b}_0$ satisfies the following constraint:

$$\vec{b}_0 = \arg\min_{\vec{b} \in B} \| \vec{a} - \vec{b} \|$$  \hspace{1cm} (3.9)

The similarity difference for vector $\vec{a}$ corresponding to the feature set $B$ is calculated as

$$f(\vec{a}) = \| \vec{a} - \vec{b}_0 \|_1$$  \hspace{1cm} (3.10)

where $\| \cdot \|_1$ represents the absolute difference of feature vectors. $f(\vec{a})$ has the same dimension as feature vector $\vec{a}$ and $\vec{b}$. In this work, the dimension of $f(\vec{a})$ is 128. The similarity from feature vector $\vec{a}$ to $B$ is calculated by

$$\tilde{h}_{vec}(A, B) = \frac{1}{n} \sum_{\vec{a} \in A} f(\vec{a})$$  \hspace{1cm} (3.11)

This represents the vector average over all $n$ interesting points in $A$ with their minimum distances to set $B$. Symmetrically, $B$ to $A$ is also computed using the same scheme described above. The final MHD is calculated by

$$\tilde{H}(A, B) = \max_c(\tilde{h}_{vec}(A, B), \tilde{h}_{vec}(B, A))$$  \hspace{1cm} (3.12)
where $\text{max}_c$ represents the component-wise max operation, and the output is a vector with the same dimension as the feature operator. After calculating the facial similarity MHD of all four facial parts, we arrange them into a long vector $\vec{H}$, where $\vec{H} = \{\vec{H}_{re}, \vec{H}_{le}, \vec{H}_{nose}, \vec{H}_{mouth}\}$, where $re$ and $le$ represent right and left eye, respectively. $\vec{H}$ represents the similarity of one face pair within a photo. There are $C^2_L$ face pairs to compare for a photo with $L$ faces. $\vec{H}$ is used as the Pairwise Feature (PF) to represent the facial similarity between two face images.

**Deep Neural Network Extraction:** Our another solution is to adapt deep neural network to extract facial feature. One major advantage of the deep neural network is its improved capacity in representing face characteristics, especially compared to previous hand-crafted features (LBP [53], SIFT [16], etc). This has been proved in previous works for face recognition [54, 55, 56]. Experimental results reported in these works have demonstrated that Convolutional Neural Network (CNN) [57] can greatly improve the face verification performance. Especially the obtained performance on popular benchmarks (the Labeled Faces in the Wild (LFW) [58] and Youtube Faces (YTF) [59]) has outperformed the human being. There are several reasons why convolutional neural network can help obtain very good performance. One reason is because of the availability of large dataset. Another reason is because of the availability of scalable computation resources, especially GPU technology.

Motivated by recent successful applications of deep neural network in face identification and related works in measuring the similarities between people with kinship relation, we attempt to apply deep models in extracting appearance features in group photo. Specifically, we use CNN as the basis to extract middle-level appearance feature to characterize facial similarities of people in the group photo. Meanwhile, due to the influence of age in measuring the facial similarity [60, 61], especially the age’s influence in kinship verification [14], we also incorporate the age information in our proposed appearance model. The illustration of the whole framework is illustrated in Figure 3.6.
**Figure 3.6:** An illustration of the proposed DOGF feature extraction. To be concise, only one face pair is listed here. Other facial pairs all follow the same processing procedure. In the codebook generation step, two colors represent different classes (family and non-family), which are represented by green and blue. In each category, four codebooks are learned from corresponding pairwise feature set divided by different age gaps.

All aligned face images are resized into the same size of $227 \times 227$ with three channels (RGB). Our basic CNN structure includes five convolutional layers following by two fully connected layers and a softmax layer. Feature maps of the first convolutional layer are calculated by the convolution between the input RGB image and 96 different kernels with a stride of 4. We set the size of these 96 kernels as $11 \times 11 \times 3$. The size of sequential convolutional layer filters are $5 \times 5 \times 256$, $3 \times 3 \times 384$, $3 \times 3 \times 384$ and $3 \times 3 \times 256$. The output neurons obtained from two fully connected layers are 4096 and 530 respectively. ReLU function [62] is applied as the activation function of all convolution layers. Max-pooling layers are applied after each convolutional layer. The size of block in max-pooling layer is set as $3 \times 3$ with a stride of 2. To train the whole framework, we use back-propagation along
with the loss computed in the softmax layer as indicated in Figure 3.4.

We use Caffe [63] as the toolbox to build the proposed deep neural network. To initialize the weights, a Gaussian distribution with zero mean and a standard deviation of 0.01 is used. We use zeros as the initial value for the biases. We set the batch size of 128. The momentum is set as 0.9 and the weight decay is set as 0.005 for all layers. We use the FaceScrub dataset [33] to train the whole network.

We use feature maps extracted from the second fully connected layer to represent the facial characteristics. Assuming the representation for two facial images are \( A = \{ \vec{a}_1^a, \cdots, \vec{a}_e^a \} \) and \( B = \{ \vec{b}_1^b, \cdots, \vec{b}_e^b \} \), where \( e \) is the number of feature dimension. \( PF = \| A - B \|_1 \) is calculated as the pairwise feature (PF) to measure the facial similarity between a pair of face images. For a group photo with \( n \) different individuals, there are \( J = C_n^2 \) different face pairs in total. The pairwise feature set of a group photo including \( n \) faces can be represented as \( PF_{\text{photo}} = \{ PF_1, PF_2, \ldots, PF_J \} \).

### 3.4.2.2 Facial Codebook Construction

A given photo usually has many different kinship relations. Age progression usually restricts the performance of measuring the facial similarity. In this framework, we integrate the extracted pairwise feature and age label information together to build a codebook scheme for measuring the facial appearance. We use seven different age categories in the whole framework. These age categories include \([1, 5, 10, 16, 28, 51, 75]\). These age labels indicate infant, kid, school-age child, teenager, youth, middle-aged adult and elder. Since our appearance model is used to measure the similarity of facial pairs, we calculate the age gap between two compared individuals. We categorized four different age gaps as indicated by \( G_{\text{age}} < 10, 10 \leq G_{\text{age}} \leq 20, 20 < G_{\text{age}} \leq 40, G_{\text{age}} > 40 \). Based on the age gap division, we divide the whole pairwise feature set \( PF_{\text{photo}} \) into four different groups. Within each pairwise
feature set, we apply K-means \cite{64} to learn facial similarity codebooks. There are \( H \) different codewords in each codebook. \( H \) is set to 9 in this work. We have two different categories, which are family and non-family. For each category, codebooks are learned separately. Two different codebook sets are represented as \( \hat{G}_{mh} \) and \( \hat{G}'_{mh} \)(\( h = 1, 2, ..., H \))(\( m = 1, 2, ..., M \)), where \( m \) indicates the number of age groups and \( h \) represents the number of codebooks. In this work, \( M = 4 \). For different group photos with different relations, age gaps calculated from different face pairs can vary a lot, construction of codebook groups based on age information is therefore necessary.

3.4.2.3 DOGF Feature Extraction

We calculate pairwise feature \( PF_j \) for \( j \)th face pair in the group photo and estimate the specified age gap group \( m \). After we estimate the age gap between two faces, we can find its two corresponding codebooks, \( \hat{G}_{mh} \) and \( \hat{G}'_{mh} \). Then we compute the pairwise cosine similarities between \( PF_j \) and \( (\hat{G}_{mh}, \hat{G}'_{mh}) \). The computation result can be represented as \( \vec{d}_j = (d(PF_j, \hat{G}_{mh}), d(PF_j, \hat{G}'_{mh})) \). Since age codebooks used for each facial pair with different kinship relations are not the same, the calculated similarity feature using our proposed scheme can also capture the co-occurrence of different relations in different photos.

For a group photo including \( n \) different individuals, we can extract \( C^2_n \) different pairwise cosine similarity facial features. We name the appearance feature for representing facial similarities in the given group photo as Degree Of Group similarity Feature (DOGF). DOGF is the feature descriptor to measure facial similarities of people in group photo. To represent the facial similarities of one group photo, DOGF can be calculated as \( \overrightarrow{F} = \frac{1}{J} \sum_{j=1}^{J} \vec{d}_j \), where \( J = C^2_n \). \( \overrightarrow{F} \) is directly used as appearance feature for photo categorization in this work. In the testing phase, we estimate the age of one given facial image using scheme proposed in [65].
3.4.3 Semantic Model

From the illustration of group photos, we observe that the context of where a picture is taken can also be used to extract valid cues to predict the category of the group photo. Intuitively, the places where people go and the activities they perform tend to be different when in the company of family members as opposed to with their friends. For example, it is more likely for people to go to nightclub with friends rather than family members, while dinners at someone’s house would seem more likely to be family gatherings. Though this observation, we investigate whether the semantics of the context in which pictures are taken hold any correlation with type of group people in the picture (family or not family). To validate our idea, we trained a set of 764 visual semantic models from a set of half a million images downloaded from the web, following the framework introduced by the IBM Multimedia Analysis and Retrieval System (IMARS) [66]. The training dataset were manually labeled and organized in a hierarchical faceted taxonomy. These categories are related to “objects”, “scenes”, “people”, “activities” and “events”. Each model $SC_i$ is an ensemble of SVMs with linearly approximated $\chi^2$ kernel. It is learned on top of bags of examples randomly sampled from the set of manually labeled web images. The features used to trains SVM model include several different visual descriptors, such as the color histogram, color correlogram, wavelet texture, edge histogram, gist, LBP histogram, etc. These features are extracted from multiple different regions of the image in a similar fashion to the spatial pyramid framework.

The score of a Semantic Concept $i$ on in a new image $x$ is then calculated as

$$SC_i(x) = \sum_{k=1}^{N_i} w_k b_k(x).$$

(3.13)

$SC_i(x)$ is the weighted sum of the scores on $x$ of the individual SVMs, which we define base models $b_k$ in the ensemble. The weights $w_k$ are learned via cross-validation during training.
Finally, we normalize the score into the \([0, 1]\) range by fitting a sigmoid on the prediction scores of the validation set.

In our problem, we map each photo \(x\) to the semantic space by concatenating all the models scores into a \(N\)-dimensional Semantic Model Vector [67], that is, a vector in which each dimension has a semantic meaning corresponding to the prediction of a visual classifier to the given picture

\[
SMV(x) = [SC_1(x), ..., SC_i(x), ..., SC_N(x)].
\]

We then use this concatenated vector as a feature on top of our trained SVM model to distinguish family from non-family pictures, as explained in the following Section. Our experimental results described in Section 3.5 illustrate that the visual semantics of a picture, even if not as powerful as other features, indeed can be used to predict whether a picture belongs to the family category or not. Furthermore, it proves to be a complementary cue which can be integrated with other descriptors to improve recognition performance.

In order to qualitatively evaluate our selection of visual classifiers and determine the most discriminative ones for family pictures, we trained two linear SVMs on top of the Semantic Model Vector representation: one using family pictures as positives and non-family ones as negatives, and the other inverting the roles. Figure 3.7 reported the top five weights of the SMVs. We use blue to denote family category, whereas, red is used to represent non-family class. The larger the weight, the higher the association of a visual concept with a category of pictures versus the other. For each category, we list the top three and bottom three pictures recognized using the semantic model. The rank is based on their visual classifier scores (shown below each photo). However, the visual classifiers are not usually perfect and sometimes can make errors. We highlight them using a red border in the Figure 3.7. We also find that even when the exact semantic information is lost, the classifier still has
the discriminative power to contribute to the picture classification. For instance, the first three images listed in the first row do not include a boat, but the classifier is picking up a correlation between family photos and setting close to bodies of water. It is interesting to observe that people tend to eat out at restaurants and go to sporting events with friends rather than family. Even the rooms of the house can be an indicator of who we are taking pictures with: in the bedroom with our family, in the living room with friends.

3.4.4 Classification

Our goal is to classify a given group photo into two different classes, family or non-family photo. This topic can be formulated as a binary classification problem. In our system, we use SVMs [68] with RBF kernel as the classification method. In the first step, each model is trained on top of each feature separately. To get the best parameter, we follow the optimization scheme advocated in [30] where a five-fold cross validation approach and grid search over the parameters space are used to estimate $C$ and $\gamma$ for the RBF kernel. In this system, there are three different features, each feature can be considered as a weak learner towards the final task. To fuse different features together, we can apply many different approaches, such as feature level early fusion and score-level late fusion [69, 70]. In this work, we use weighted fusion scheme on the output of each RBF kernel obtained in each individual model. The fusion is formulated as follows:

$$R_c(x, y) = \sum_m w_m e^{-\frac{|x-y|^2}{2\sigma_m^2}}, \quad \sum_m w_m = 1,$$

where $R_c(x, y)$ is the combined kernel value for samples $x$ and $y$, and $\sigma_m$ is the RBF parameter of kernel $m$. $w_m$ is the weight associated with $m$ th model (appearance, geometry, semantic information). In this work, $w_m$ is obtained via cross-validation on the training
data. We follow the scheme advocated by Ayache et al. [71]. \( x, y \) are the feature vectors associated with the model. In our problem, there are three different models.

3.5 Experiments

In this section, we introduce the details of the used dataset and illustrate the performance obtained by different models. Following this, we list the experimental results and give analysis. In order to provide a deeper understanding of the proposed model, we also demonstrate classification results obtained from different models individually, e.g., geometry, appearance and semantic model. In addition, we also give the comparison between our scheme with previous works [1] under the same experimental setting.

3.5.1 Dataset

The first work for family photo classification is proposed by Chen et al. [1]. They used a dataset including 1,167 family photos and 1,263 non-family photos. To balance the number of samples in positive and negative categories, we add more photos to each category. The expanded dataset that we are using includes 1,420 samples for every category (family and non-family). In order to collect the dataset, we followed the same collection scheme as Chen et al. [1]. The images came from a public dataset collected by Gallagher and Chen [34]. The images included in the dataset are collected from social media (e.g., Flickr) using keywords, such as “family portrait” and “group photo”. This dataset provides some initial labels, such as family, group, wedding. However, from the perspective of our classification problem of interest, the dataset presented some labeling errors. For example, some family photos appeared in the group category, and several non-family photos were included in the non-family photo category. We corrected such mislabelings in our newly extended and organized dataset. Five people were involved in labeling the new extended dataset. A photo was labeled as family photo or non-family photo only if all members agreed on its label. Otherwise the photo was not used in the experiments. In our work,

38
Table 3.1: Classification results using different schemes. Fusion Scheme I indicates the fusion scheme based on geometry and appearance information (CNN). Fusion Scheme II indicates the fusion model using appearance, geometry and semantic information together. Fusion Scheme III indicates the fusion scheme based on geometry and appearance information (dense SIFT based). Chen et al.* indicates the assumption that approach [1] could classify the newly added images on the collected dataset with 100% classification accuracy.

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chen et al.*[1]</td>
<td>91.7%</td>
</tr>
<tr>
<td>Geometry (Ours)</td>
<td>87.3%</td>
</tr>
<tr>
<td>Appearance via SIFT(Ours)</td>
<td>89.0%</td>
</tr>
<tr>
<td>Appearance via CNN(Ours)</td>
<td>90.2%</td>
</tr>
<tr>
<td>Semantic (Ours)</td>
<td>76.3%</td>
</tr>
<tr>
<td>Fusion Scheme I (Ours)</td>
<td>95.1%</td>
</tr>
<tr>
<td>Fusion Scheme II (Ours)</td>
<td>96.7%</td>
</tr>
<tr>
<td>Fusion Scheme III(Ours)</td>
<td>93.4%</td>
</tr>
</tbody>
</table>

we don’t consider group photos where family members and friends are mixed. As illustrated in Figure 3.1, the collected dataset presents a wide variety of subjects in different poses including sitting, standing or laying. From the illustration, we can also find that it is a very challenging dataset for measuring facial similarities because of many variations in faces, such as occlusions, changes of facial expressions, etc.

Meanwhile, the group photos used in the experiment all include three or more people. For these photos with two people, standard kinship verification approaches [6, 11, 12] can be used to predict the pairwise relation. Our work aims at determining the category of group photos. This is different from the general kinship verification problem. This is also different from the work proposed by Fang et al. [31] where a corresponding family is predicted when given one probe face image. In general, our work is to recognize the category of group photos (family or non-family) instead of individual facial images.
3.5.2 Experimental Results

We use five-fold cross-validation to evaluate the performance of our system. For each fold, the accuracy is calculated as

\[
\text{Accuracy} = \frac{N_{\text{correct}}}{N_{\text{total}}} \times 100\% ,
\]  

(3.15)

where \( N_{\text{correct}} \) indicates the number of correctly classified photos and \( N_{\text{total}} \) is the total number of samples in the testing set. The final performance is computed by averaging the accuracy obtained all five folds. Our experimental results are illustrated in Table 3.1. From the result, we can find that each separate model performs well. This also demonstrate that the discriminative pattern between two different categories truly exists and our geometry model can capture these discriminatory information. Our proposed appearance feature DOGF also works well in the unconstrained dataset. Although the image samples used in our experiment are collected from unconstrained environments, the obtained performance is still quite promising. The appearance model alone (CNN based) can achieve a classification accuracy up to 90.2\%. This has improved the performance compared to the scheme which uses SIFT feature with Modified Hausdorff measure scheme. Part of the reason is because of the efficiency of supervised deep learning model in representing facial characteristics. Although contextual semantic information by itself does not achieve a comparable performance compared to the geometry and appearance model, it provides complimentary information. Experimental results obtained by Fusion Scheme II demonstrate that fusing with semantic information can help boost the classification performance. The final accuracy after fusing all different models is 96.7\%. Because Chen et al. [1] does not public the source code, we can’t directly compared with their scheme on our collected dataset. However, we assume that their approach can classify the images added to the newly collected dataset with 100\% accuracy. Under such assumption, their model would achieve a classification accuracy 91.7\%,
still almost 5.0% lower than our proposed framework.

3.5.3 Discussion

As analyzed above, our geometry feature extraction is purely based on the position pattern of the people in the group. It still gives very good performance in the group photo categorization, especially it performs very well in such family photos where people stand in a typical pattern, e.g., parents stand in the back of the children and the elders stand in the center. However, our geometry model does not work well in situations where individuals are positioned in an atypical way, e.g., standing in a row as illustrated in Figure 3.8(b).

To measure the performance of the advocated geometry model, we also compare the proposed geometry feature with another baseline where we calculate the distance between the centroid and all vertices lying on the polygon. We then get the mean value of all these distances and use the mean distance as the feature for classification. The comparison result is listed in Figure 3.12(a) and demonstrates that the proposed geometry feature performs much better than the baseline in characterizing the standing pattern of the group photo.

The experimental results obtained by DOGF feature are listed in Figure 3.9(a) and Table 3.1. We can find that the proposed DOGF feature is very efficient in discriminating two different categories. Our appearance model can also compensate with failure cases produced by the geometry model where family members stand in a less traditional way, such as standing in a row. It also works very well in other difficult situations, for example, when people sit around the table (e.g., dinner). Meanwhile, in cases where the appearance feature does not work, we can rely on the geometry model. Our experimental results demonstrate that these two cues are complementary to each other.

For appearance model, we also compare the performance using age information vs without using age information. As illustrated in Figure 3.12(a), experimental results show
that age information plays an important role in the final classification result. Without employing age information, the performance is very low. This result is consistent with the findings of previous studies [1]. These results show that the DOGF feature is not only very efficient but also quite effective in measuring the facial similarity in group photo.

As demonstrated with examples in Figures 3.7 and 3.10, we find that the semantic information along with the context (places, objects, people and activities) in which the two categories of photos are taken provides some insights on whether or not a picture is a family photo. Family photos usually include larger groups and are taken inside homes or, when outdoors, more likely within natural environments. On the other hand, we find that pictures belonging to the non-family category are more likely to be at sporting events or restaurants. From experimental results, we can observe that while the visual semantic information by itself constitutes the weakest cue to distinguish family photos from other ones, it also provides complimentary information with respect to other features. This can help improve the recognition performance if integrated in an appropriate fusion scheme.

3.6 Conclusion

I have developed a novel framework to automatically classify family photos and non-family photos. My work introduces multiple contributions: First, I have advocated a novel geometry feature to characterize the social relationship in a group photo. Second, a face descriptor extracted from a deep neural network architecture is applied to measure similarities of individuals in a group photo with the goal of predicting their relationships. Third, semantic information about the context in which a picture was taken is incorporated into our model to further improve the recognition performance. In the end, we have combined multiple cues in a fusion scheme. The fusion scheme increases the recognition performance by more than 6% compared to each single model, demonstrating that the proposed features can complement to each other.
Figure 3.7: Top five most discriminative semantic concepts associated with family (top) and non-family (bottom) photos. Under each photo we report the score of the corresponding semantic model. Images with red border represent mislabelings of the classifier.
Figure 3.8: (a) Illustrations of correctly classified photos using the geometry model. (b) Examples of misclassified photos using geometry model.
Figure 3.9: (a) Illustrations of correctly classified group photos based on appearance model (DOGF). (b) Illustrations of the photos misclassified using geometry model and correctly classified by the fusion model based on geometry and appearance cues.
Figure 3.10: Illustrations of the photos correctly classified using the semantic model.
Figure 3.11: Performance comparison of different models. Fusion Scheme I indicates the fusion model using geometry and appearance information. Fusion scheme II indicates the fusion model using appearance, geometry and semantic information together.
Figure 3.12: (a) Performance comparison between mean distance and the proposed geometric feature scheme. Mean distance indicates that we calculate the center position of the polygon, then calculate the distance between the center to all the vertices in the polygon. Afterwards, the calculated mean distance is used as the feature to classify the photo category. (b) Performance comparison for different appearance models. No age indicates that there is no age information used in appearance model.
4.1 Background and Related Works

Facial attributes play significant roles in many applications in human-computer interaction (HCI) and artificial intelligence systems. To evaluate the performance of these systems, one important criterion is to determine if they can correctly interpret facial images in real time. Acquiring these facial attributes accurately from images greatly improves these systems’ performance. For example, facial image analysis can help verify the identity [72, 73, 74], gender [75], age [76, 77], expression [78, 79], ethnicity, etc. It even can help identify the kinship relation between people [11, 80].

Among these facial attributes, age information is considered as one of the most important soft biometric traits for human identification or search. Age encapsulates key demographic information. Compared to other recognition problems (e.g., object classification, scene categorization) in computer vision, age estimation is much more challenging since the difference between facial images with age variations can be more subtle and the aging process varies greatly among different individuals.

Automatic age estimation systems have wide applications. Age information can be used in law enforcement where it can be used to help reduce the search range of the gallery set by locating the suspects by locating the suspect’s age within a specific age group. This can improve the efficiency of matching. Age estimation systems also have many useful applications in commercial domain, for instance, store owners can adjust decoration styles and advertisements according to their customers’ demographics. It can also be used in other
Figure 4.1: An illustration of the aging process of two different individuals. Each row are images of the same person at different ages.

areas as well, such as a well designed Age Specific Human-Computer Interaction system (ASHCI) can help prevent minors browsing adult web pages or purchasing age restricted material from vending machine. ASHCI can facilitate our daily life, for example, an ASHCI installed vehicle can prevent children from starting the engine without the guidance of adults. An ASHCI system can give a warning when the driver leaves the kids alone in the vehicle without taking any protective measures.

Recently, facial attributes have been widely used as a soft biometric for identification [81]. This work demonstrates that utilizing facial attributes can help improve the recognition performance. Age information is not only used in face recognition problem, but also in many other topics. In kinship verification [14], xia et al. found that age variance between family members affect the verification performance greatly. Age information was used as the priori-knowledge for training the kinship verification classifier. Wang et al. [32] utilized age information as a feature to predict the social relation between people in a given photo. In [1],
Chen et al. combined age information with other attributes to construct the graph as a feature for classification, which has been proven to be effective in family photo classification.

Although facial image age estimation is an important technique in real-world applications, it is still very challenging. As demonstrated in Figure 6.2, we can find that the aging process of different individuals varies greatly. The aging process is very complicated as demonstrated in [82, 83]. There are influential factors affecting the aging process. In general, these factors can be divided into – internal and external categories. The internal factors are determined by physiological factors, such as genes. External factors contain many elements, such as living environment, health condition, lifestyle, etc. Even air pollution, stress and sun exposure can affect facial aging progress. Extracting robust aging features which are invariant to the influence due to individual differences is still an open problem. To accurately estimate age, in this work, we explore a new framework for image based human age estimation problem.

4.2 Related Work

Age estimation from facial images has been explored by many researchers over the last few decades [84, 85, 86, 65, 87, 88, 89, 90, 91]. In general, the age estimation problem is composed of two different sub-problems: aging feature representation of facial images and prediction of the age using the extracted feature. For the aging feature representation, there are several representative works. One work was conducted by Kwon and Lobo [84]. Their method was based on the cranio-facial development theory and analysis of skin wrinkles. Anthropometric information was used to predict age for the given face. Ramanathan and Chellappa [92] adopted the similar idea and proposed a craniofacial growth model to characterize the facial shape growth variations, where facial landmarks were used to calculate the growth parameter. Their system was designed for coarse age estimation, such as classifying a probe image into a general age group - babies, young adults, or seniors [84]. One limitation
of this kind of work is that shape information is not accurate at predicting the exact age.

Another representative work is using texture information to characterize aging cues. Biologically inspired features (BIF) designed for object classification [85] were introduced to age estimation problem by Guo et al. [65], where a set of predefined Gabor filters with different scales and orientations were used as filters to extract aging features. BIF has been proven to be one of the most efficient aging feature [93] and has been used in many related works [94, 95, 96]. In view of the success of BIF in age estimation, many works advocates to optimize the BIF feature extraction scheme. One representative work is proposed by Chang and Chen [97]. They advocated an aging descriptor using scattering transform on the basis of traditional BIF extraction framework, where Gabor coefficients are scattered and pooled with Gaussian smoothing in layers. Feature fusion is also applied in this area, one representative work is conducted by Ren and Li [91] where three different manually crafted features (Gabor wavelets, HOG and SIFT) are integrated together as the aging representation to predict age. Their results demonstrated that feature fusion performs better than using each separately. To capture facial local information, Lu et al. [98] advocated a cost-sensitive local binary feature learning (CS-LBFL), which is based on the projection of raw pixel values extracted from the local facial patches via a series of hashing functions. With the popularity of Convolutional Neural Network (CNN), recently, Yi et al. [93] and Wang et al. [99] applied CNN into the age estimation scheme. Their experimental results also demonstrate that deep neural network can be applied in age estimation problem. Meanwhile, these two works found that designing a deeper neural network may not help improve the estimation performance. One possible reason is because of the aging information is very subtle to capture compared to other topics, such as object classification.

Subspace analysis is also applied in the age estimation problem. Geng et al. applied aging pattern subspace to represent aging information from images [86]. In general, their
idea was to construct the aging subspace based on the sorted facial images for each individual. Their experiments showed that the proposed scheme can deal with missing ages in the image sequence. Fu and Huang [87] found that aging patterns extracted from the projected subspace by manifold learning were discriminative and effective at age estimation. The results obtained in their work showed that manifold mapping algorithms were effective at age estimation problem, as it can greatly reduce the redundancy of original image space while providing effective discriminant representation for regression. Local manifold structure was also applied [89]. Their experimental results also illustrated that preserving local manifold structure can improve the performance of age estimation. A ranking scheme was proposed for dealing with the age estimation problem [88, 89, 97]. Their schemes were motivated by the ordinal characteristics of aging process, such as a 9 year old’s face is much more closely related to the face of an 11 year other than a 40 year old.

The age estimation problem can be viewed as a classification [86] or regression problem [100, 101, 87]. Recently, several new schemes are introduced to the age estimation problem. Partial Least Squares (PLS) regression and Canonical Correlation Analysis (CCA) were used to learn the aging pattern joint with gender and ethnicity [102, 103]. To deal with sparse and imbalanced training data as well as variations of feature caused by different intrinsic ambiguities, Chen et al. [104] used cumulative attributes to learn the regression model for age predication. Han et al. [96] proposed a hierarchical estimation framework which consists of inter-group classification and intra-group regression for human facial image age estimation scheme. To minimize the influence of low-quality image to the age estimation problem, quality assessment scheme was also introduced in their work. Furthermore, they also compared the performance with human perception in age estimation problem.

In this chapter, we explore learning aging features directly from the data. The learned aging features are expected to be more general and discriminative than previous pre-defined feature extraction algorithms [86, 83, 105]. Our approach has several novelties. First, we
propose a novel method for automatic age estimation. The proposed scheme is based on a deep learning model (CNN). The whole scheme is supervised. Second, to evaluate the performance of the proposed approach, we compare with the state-of-the-art approaches on two benchmark datasets. Experimental results show a significant improvement compared to these state-of-the-arts. Third, we adopt manifold learning approaches with this scheme. An evaluation of different manifold learning approaches is also conducted. Furthermore, different regression and classification approaches are evaluated based on the deep learned aging features obtained from the proposed network.

Below is the summary of our proposed pipeline. The general structure of the framework is illustrated in Section 5.2. We give a detail description about the proposed scheme. Different manifold learning schemes are also analyzed. In Section 5.3, we talk about the dataset and experimental setting. Experimental evaluations of the our approach and the comparison with previous methods are discussed in Section 5.3. The final conclusion is given in Section 5.4.

4.3 Methods

In the following, we first give the introduction of CNNs scheme, then we talk about feature extraction scheme. A brief introduction of Convolutional Neural Networks is given, mainly focusing on the reason why the learned “deep” feature are effective in characterizing age patterns and the way to make use of the learned feature in age estimation problem.
Afterwards, we study different manifold learning approaches in age estimation for its good performance to capture the underlying aging structure and representing age information in a low dimensionality. We also investigate several general classification and regression methods including Support Vector Regression (SVR) [106], Support Vector Machines (SVMs) [36], PLS [107] and CCA [108] to estimate age.

4.3.1 Convolutional Neural Networks (CNNs)

To extract aging patterns from various facial images, we need a powerful learning model. Recently, deep learning models have been used to extract features in a hierarchical structure. The constructed structure aims to extract hierarchical, abstract, and translation invariant features. Deep learning models have already shown very promising and plausible results in many applications [109, 110, 111, 112, 113]. From these works, we find that deep learning schemes demonstrate very promising features at characterizing high-level abstractions of visual data by using a deep architecture. In general, the deep learning architecture is composed of multiple non-linear transformations. Among these deep learning models, Convolutional Neural Networks (CNNs) [57] have demonstrated an excellent ability in capturing image characteristics [111, 112, 113]. CNNs can be classified as a category of biologically-inspired, multi-layered neural networks. It is usually constructed by feed-forward neural networks. The construction of CNNs was inspired by the study in biology area where biological processes can be used by multi-layer transformations to simulate the human perception. The learned neurons captured in the feature map correspond to overlapping regions in the visual field.

CNNs has obtained very promising results in many computer vision topics recently, such as in visual object recognition, detection and image retrieval [113, 114, 115, 116]. The feature learned using CNN has demonstrated very effective in representing image characteristics. Specifically, based on the visualization of learned features illustrated in
we can find these learned interpretable features at different level of CNN layers are very efficient. Due to its good ability in extracting features from images, in this work, we use CNNs as the basis to extract aging feature from static facial images. Our experimental results demonstrate that the learned weights from the deeper layers are more specific to the age estimation problem. In general, our CNN structure is composed of convolution layer and sub-sampling (pooling) layer which are stacked in alternative. A fully connected MLP (multiple logistic perception) is applied at the top of the whole framework.

4.3.1.1 Convolution Layers

In convolution layer, each neuron is formed by the input from a local receptive field in the preceding layer and the learned kernels (weights). Neurons within the same feature map share the same kernels but from different input receptive fields. The kernels used for different feature maps in the same layer are different. Subsequently, an activation function is applied. The calculation can be represented as:

\[
  u^\ell_j = \theta \left( \sum_{i=1}^{N} u^{\ell-1}_i \ast w_{ji} + b^\ell_j \right),
\]

where \( u^{\ell-1}_i \) is the input neuron from \( \ell - 1 \) layer, \( N \) is the total number of input neurons, \( b^\ell_j \) denotes the bias, \( \theta(\cdot) \) is the activation function. \( \ast \) denotes the convolution calculation. In our framework, logistic (sigmoid) function is applied as the activation function.

4.3.1.2 Sub-sampling Layers

Sub-sampling layer down-samples input feature maps. The calculation only changes the size of the input maps while the number of input maps keep unchanged. There are several schemes for sub-sampling operations, such as averaging or taking the maximum, or using learned combinations of the neurons in the block. In this framework, average pooling is applied.
The purpose of sub-sampling operation in the framework is to maintain specificity. Although the concept is easy and simple, it is efficient in characterizing the feature for many topics. This concept adopts the general idea from the mammalian visual cortex [117, 118]. Its learned architecture has proven to be efficient in representing hierarchical features for these specific problems.

4.3.1.3 Training Process

As we know, the goal of CNN training is to minimize the error function. For a specific classification problem, such as \(N\) training samples belonging to \(S\) classes, the reconstruction error (sensitivity) is calculated by

\[
E = \frac{1}{2} \sum_{n=1}^{N} \sum_{k=1}^{K} (s_k^n - y_k^n)^2, \tag{4.2}
\]

where \(s_k^n\) indicates the \(k\)-th dimension of the \(n\)-th sample’s label, \(y_k^n\) is the output value of the \(k\)-th output layer unit corresponding to the \(n\)-th input pattern. From Eqn. 4.2, we can find that the error of the whole dataset is calculated by summing over all the individual errors on each neuron pattern. Thus for a single pattern, the error can be calculated as \(E^n = \frac{1}{2} \sum_{k=1}^{K} (s_k^n - y_k^n)^2\). Training stage includes feedforward pass and backpropagation pass. The feedforward pass is to transform feature maps from one layer to the successive layer with the learned parameters (weights and basis) using pre-defined activation functions. For instance, we can define the output of layer \(l\) as \(v^l = f(u^l)\), where \(v^l = W^lv^{l-1} + b^l\). In the backpropagation pass, weight \(W^l\) and \(b^l\) are updated via stochastic gradient descent. This is trained on the backpropagating the derivative of the loss (errors) obtained in the feedforward pass with respect to the parameters throughout the whole network. Before training the whole network, normalizing the data to have mean 0 and variance 1 in the feature space can usually accelerate the convergence [119]. More mathematical theory and specific details can be found in [120].
4.3.2 Deep Learned Aging Pattern (DLA)

In general, our proposed human age estimation pipeline contains several steps. This is illustrated in Figure 4.2. In this work, CNN is trained for a multi-classes age estimation task. The goal of a leaned structure is to predict the age given a static facial image. We name the feature extracted from the proposed scheme as deep learned aging pattern (DLA).

As illustrated in Figure 4.2, the whole framework is comprised of 6 layers. In the beginning, we first align all the facial images based on the eyes’ coordinate, which make them lie on the same coordinate. Then we resize all the images to the same size of 60 by 60 pixels. In this work, grayscale image is used as input to the first convolutional layer with the kernel size of $5 \times 5$. As we know, after applying the filter with the size $a \times b$ to a feature map of $h \times w$, the size of the output is $(h - a + 1) \times (h - b + 1)$. As a result, the size of feature maps in layer $L_2$ after convolution is $56 \times 56$. At the end of the layer, an activation function (logistic function) is applied to obtain output feature maps. For the units within a feature map, they all share the same set of weights from the filter. Then, the obtained feature maps in the convolution layer go through a max-pooling layer.

The $L_3$ layer is computed by max-pooling the feature map obtained in layer $L_2$. $2 \times 2$ is used as the window size for subsampling in each feature map in the $L_2$ layer. This can reduce the spatial resolution while keeping the number of feature maps unchanged. As discussed in [113], max-pooling layers are used to “filter” the convolution networks’ output. This operation can deal with the local translation and rotations. The next convolution layer $L_4$ is obtained by the convolution between filters with kernel size of $7 \times 7$ and the feature maps from layer $L_3$. The next layer $L_5$ is also obtained by the max-pooling operation of feature maps from layer $L_4$. The $L_6$ layer consists of 80 feature maps of size $1 \times 1$. Each unit is connected to all feature maps in layer $L_5$. The output layer (softmax classifier) is fully connected to $L_6$ layer. Feature maps obtained in these layers are constructed to extract low-level features, including edges and texture. During the training stage, face images with
age labels are used for the supervised learning. Based on these “learned” filters, we can get the image representation which can characterize the human facial image age information. A neighborhood in one layer is connected to the units in the successive layer. This setting adopts the idea of extracting feature from local receptive fields. Another advantage of applying CNNs in this work is filter weights are shared by the units in the same feature map, which can greatly reduce the number of parameters for training.

For these works using a deep learning model based on convolutional neural network, the output of the top layer is usually used as the feature representation in their problem [121, 122]. However, motivated by the idea in [116], unlike these works, where representations in different layers respond to particular activations, in this framework, we investigate the use of extracted features in the problem of age estimation and to explore whether combing these features is an effective way for predicting age. As far as we know, this is the first time that deep learned features are extracted in such a way for age estimation.

After training the CNN, we extract features learned from different layers. We can see that the dimensionality of features in CNN layers is high, such as layer $L_4$ has a dimension as high as 5808 by itself. With the consideration of the simplicity and efficiency, principal component analysis (PCA) [123] is applied for feature dimension reduction. Assuming $X_l = [x_1, \ldots, x_N]$ to be the feature vectors of $N$ samples with the dimension $T$ extracted from the $l$-th layer. PCA aims to find the projection matrix $P$, which projects $X_l$ into the new subspace $Y_l$ based on $Y_l = P^T X_l$, where $Y_l = [y_1, \ldots, y_N]$ with dimension $t$, and $t < T$. $P$ satisfies

$$P = \arg \max_P P^T C P,$$

where $P^T P = I$. $C$ is the covariance matrix, $C = \sum_{j=1}^{N} (x_i - \bar{x})(x_i - \bar{x})^T$, $\bar{x}$ is the mean vector obtained from training samples. From the experimental results, we found that the age estimation error almost does not change when using PCA as the dimension reduction.
algorithm. This operation is very efficient in speeding up training of the age estimation model.

Afterwards, features extracted from different layers are concatenated together to obtain the aging pattern. This can be represented as \( F = [Y_{L_1}, \ldots, Y_{L_J}] \), where \( Y_{L_j} \) is the feature extracted from layer \( L_j \). In this work, the feature maps extracted from layer 2 to layer 5 are concatenated together as the aging feature (DLA).

Manifold learning is utilized in the proposed framework for its good performance in learning aging patterns and capturing the underlying face aging structure [87]. Given the feature vector \( F \) with the size of \( N \times d \), where \( N \) is the number of training samples and \( d \) is the number of feature dimensions. Manifold embedding aims to learn the projection matrix \( M \) with dimension \( d \times k \) which satisfies \( Y = M^T F \), where \( k < d \). In this framework, three different manifold learning schemes are analyzed and compared which are Marginal Fisher Analysis (MFA) [124], Orthogonal Locality Preserving Projections (OLPP) [125] and Locality Sensitive Discriminant Analysis (LSDA) [126].

Manifold learning is very efficient at feature dimension reduction and discriminative learning. However, one disadvantage of manifold learning is that it is very sensitive to image misalignment. In the age estimation problem, facial images span a very large age range. This makes it very difficult to align different facial images together. Large facial shape changes of different persons at various ages also makes the problem more difficult to handle. As we talked above, the deep learned feature can tolerate these translations, rotations and scale changes in images with max pooling operations in the framework [116]. From this point, we can see that application of manifold learning in deep learned feature is promising for robustness of deep learning model (insensitive to images misalignment) and discriminative power (supervised learning with the age label information).
Table 4.1: Illustration of the used datasets. \( N_{in} \) indicates the number of individuals. \( Age_r \) represents the age range of the corresponding dataset.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>( N_{in} )</th>
<th>( Age_r )</th>
</tr>
</thead>
<tbody>
<tr>
<td>MORPH [129]</td>
<td>5475</td>
<td>16-77</td>
</tr>
<tr>
<td>FG-NET[130]</td>
<td>1002</td>
<td>0-69</td>
</tr>
</tbody>
</table>

4.3.3 Regression or Classification

As talked in previous works [127, 128], age estimation can be considered as a regression problem, since each age can be considered as a regression value. Meanwhile, the age estimation problem can be categorized as a classification problem for the reason that we can also classify each age as a class label. This is the first time that a deep learned feature is extracted in this way for age estimation, it remains unknown as which scheme performs better. In our framework, both SVMs for age classification and SVR for age regression are evaluated and compared. We also use PLS and CCA as regression algorithms in this work for their good performance in the age estimation problem [103].

4.4 Experiments

In this work, we use two different measuring criteria to evaluate the performance of the proposed scheme in age estimation. They are Mean Absolute Error (MAE) and the Cumulative Score (CS). The MAE is calculated using the average of the absolute errors between the estimated ages and the ground truth (labeled age). MAE is calculated as

\[
MAE = \frac{1}{K} \sum_{k=1}^{K} |g_k - g_k'|, \tag{4.3}
\]
where $g_k$ and $g'_k$ indicate the ground truth age and estimated age respectively. $K$ is the total number of test images. The cumulative score (CS) is computed as:

$$CS_e = \frac{K'_e}{K} \times 100\%,$$

where $K'_e$ denotes the number of probe facial images whose absolute error between the estimated age and the ground truth age is not greater than $e$ years.

### 4.4.1 Datasets and Experimental Settings

In this work, to evaluate the proposed method and compare with state-of-the-art approaches, our experiments are conducted on two available public datasets, MORPH database [129] and FG-NET [130]. To compare with state-of-the-art approaches, we adopt the same experimental setting as used in [83, 104, 89]. The illustration of used datasets is given in Table 4.1.

Following the experimental setting in [88, 104], in MORPH database, we randomly divide the whole dataset into two subsets, one subset is used for training, and the other one is used for testing. The ratio between the training and testing is 4:1. It means 80% data is used for training and 20% is used for testing. There is no overlap between the training and testing sets.

In FG-NET, leave-one person-out (LOPO) setting is applied. It is same as the approaches conducted in [131, 83, 88, 104]. This is determined by the size of FG-NET. FG-NET only includes 1002 facial images belonging to 82 subjects. However, FG-NET has a wide age range which is ranging from 0 to 69 years old. This wide age range is the major reason for its popularity in age estimation work.
Table 4.2: Comparison of different methods for age estimation on two benchmarks (measured in MAE).

<table>
<thead>
<tr>
<th>Method</th>
<th>FG-NET [130]</th>
<th>MORPH [129]</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGES. [86]</td>
<td>6.77</td>
<td>8.83</td>
</tr>
<tr>
<td>RUN [131]</td>
<td>5.78</td>
<td>-</td>
</tr>
<tr>
<td>Ranking [101]</td>
<td>5.33</td>
<td>-</td>
</tr>
<tr>
<td>LARR [83]</td>
<td>5.07</td>
<td>-</td>
</tr>
<tr>
<td>SVR [83]</td>
<td>5.66</td>
<td>5.77</td>
</tr>
<tr>
<td>MTWGP [132]</td>
<td>4.83</td>
<td>6.28</td>
</tr>
<tr>
<td>OHRank [88]</td>
<td>4.85</td>
<td>5.69</td>
</tr>
<tr>
<td>PLO [89]</td>
<td>4.82</td>
<td>-</td>
</tr>
<tr>
<td>CA-SVR [104]</td>
<td>4.67</td>
<td>5.88</td>
</tr>
<tr>
<td>Proposed Scheme</td>
<td>4.25</td>
<td>4.77</td>
</tr>
</tbody>
</table>

4.4.2 Experimental Results and Analysis

The obtained experimental results are illustrated in Table 4.2 and Figure 4.3. From the results, the proposed scheme demonstrates the effectiveness and robustness in age estimation. We can find our new representation of aging feature outperforms other state-of-the-art approaches [86, 131, 101, 83, 132, 89, 104] by a large margin.

We also find that the obtained MAE is reduced by a large degree. The MAE gotten on MORPH is 4.77 compared to the current best result 5.69 achieved in [88]. The MAE from FG-NET is 4.26. It is less than 4.67 obtained in [104], which is the state-of-the-art approach. The improvement of MAE reduction is very evident. Let us use the result obtained by [83] as the baseline, our scheme improves on the baseline a full year on average $|5.77 - 4.77| = 1.00$. This improvement is more than 10 times the improvement put forth by [88]. This also demonstrates the deep learned aging feature is an effective representation for characterizing facial aging.

We can categorize the proposed schemes into three different methods into DLA+MFA, DLA+OLPP, DLA+LSDA based on the different manifold learning algorithms. In general,
discriminative mapping of the deep learned feature proved to be effective in learning the age function, especially using MFA as indicated in Figure 4.3. It performs better than OLPP and LSDA. To show the distribution of the manifold learning results and give a better understanding of the learned feature, the first three feature vectors after projection are used and the samples are displayed in Figure 4.4, which indicates that the extracted features are discriminative. In this work, three different manifold learning algorithms are used to construct a general framework for learning discriminative aging patterns. All of them have proved powerful in mapping deep learned aging pattern (DLA) into a new subspace. Even
the largest 5.51 (MAE) obtained by LSDA is smaller than 5.69 \cite{88, 104}, still an improvement compared to the state-of-the-art. The regression scheme applied in this experiment is SVR. Another experiment is conducted to validate the effectiveness of the manifold learning. Without using the manifold learning, the MAE obtained on MORPH is 5.02. This is higher than 4.77. This demonstrates the effectiveness of the manifold learning. Except these three different discriminative mapping algorithms, other manifold learning schemes can also be used. This framework could be a reference to other related work in using deep learned feature in the age estimation problem.

To analyze the performance of different classification and regression algorithms in DLA, experiments for testing performance of different regression and classification schemes are conducted in MORPH. The regression and classification algorithms include SVMs, SVR, PLS and CCA. The results are illustrated in Figure 4.5 and Figure 4.6. RBF kernel is used in SVMs and SVR, the parameters of the learners are adjusted based on the training data. For the parameters setting of PLS and CCA, we follow the introduction given in \cite{102}. From the comparison of MAEs, SVR performs best in estimating age based the extracted feature (DLA). Experimental results demonstrate that SVR performs better compared to others. The performance obtained by PLS and CCA is similar to using SVMs. One possible reason is in \cite{102}, where gender and ethnicity labels are also used in the training phase to learn the regression function. It is known that the aging progression across male and female, with different ethnicities are different \cite{133, 102, 77}. In this paper, unlike \cite{133, 102, 77}, we focus on analyzing the performance of a new aging feature (DLA) instead of considering the influence due to gender and ethnicities, the label used for training is only age information. This work demonstrates that SVR outperforms PLS and CCA when only age label is provided. The results obtained by PLS and CCA are nearly the same as indicated in Figure 4.5 and Figure 4.6.
To compare the performance between using the feature from top layer $L_5$ and the extracted feature using layers from $L_2$ to $L_5$, another experiment is also conducted on MORPH. We only use the feature extracted from layer $L_5$ as the deep learned aging feature. The MAE obtained from the top layer $L_5$ is 5.63, which is higher than the result (4.77) obtained using all the layers. But the result from the single layer $L_5$ is still better than the state-of-the-art. This also demonstrates the effectiveness of the proposed scheme over the state-of-the-art approaches. We think one reason why the extracted feature performs so well is the extracted features using CNNs is “supervised”. The supervised learning strategy in CNNs forces the network to extract the age related information at multiple different scales from images. Therefore, the learned features are more distinctive in the age information spanned manifold which can be seen in Figure 4.4.

4.5 Summary

In this chapter, I have advocated a new framework for face-image-based age estimation problem. I have introduced deep learning scheme into the problem of age estimation. We have demonstrated the effectiveness of the proposed model in predicting one’s age based on a facial image. I have evaluated the proposed system on two benchmark datasets. The results outperform the state-of-the-art by a large margin in both. We also incorporate manifold learning algorithm in the framework, showing its effectiveness in discriminative subspace learning of the deep learned age pattern. Experimental results also show that support vector regression technique performs best among the regression and classification algorithms. We have demonstrated the potential of our approach towards real world applications.
Figure 4.4: Visualization of the first three components of the extracted deep learned feature after manifold learning (MFA) on the MORPH dataset. The age labels are colorized for a better view of the feature distribution after manifold mapping.
Figure 4.5: Cumulative scores of the algorithms with different settings on MORPH dataset.
Figure 4.6: The age estimation errors using different methods on MORPH dataset.
Chapter 5
AGE ESTIMATION VIA CONVOLUTIONAL SPARSE CODING

5.1 Introduction

As an important component in facial image analysis, age estimation is gaining increasing attention. However, until recently, a social viral carnival∗ made the mobile App designers realize the true value of age estimation. We have conducted the interview with app designers and they have specified two equally-important requirements: Accurate: estimation error $\leq 5$ years; and Lightweight: the module can be executed on the mobile phone with limited memory, time and power cost. In this chapter, I show how to develop a practical age estimation module complying with both requirements.

The key to successfully estimating age is crafting a discriminative aging feature, which has been extensively studied in the literature [134]. However, manually crafting an aging feature to describe the intrinsic aging pattern is extremely challenging [135, 96]. A good aging feature should characterize various aging factors [82, 83, 135, 136, 137, 138, 86], including both internal factors (genes, gender, ethnicity [133, 138]) and external factors (lifestyle, environment, air pollution, stress, sun exposure). Among all the manually crafted features, convolutional feature map aging framework (e.g., bio-inspired feature BIF [93, 65, 133, 139, 138]) has proven to be the most successful aging representative framework. This framework leverages a set of predefined Gabor filters with different scales and orientations

∗ How-Old.net tool
to extract aging features. However, due to the fact that Gabor filters are essentially edge detectors, their convolutional feature maps are more sensitive to edges than higher order image statistics.

To overcome above drawbacks and further improve the framework’s accuracy, our insight is to adopt this convolutional feature map aging framework, but leveraging feature learning approach to automatically learn the convolutional filters instead of manual crafting. Among all the candidate learning approaches, deep convolutional neural net (CNN) has recently become the most popular feature learning approach and it dramatically improves the state-of-the-art in many facial image analysis tasks. Unfortunately, a deep CNN requires a prohibitive budget on time, memory and power consumptions during feature extraction, making the deep convolutional architecture incompatible with mobile phones. For example, in a widely used 8 layer AlexNet, the total amount of parameter weights is 61M [140], which is intolerable to app designers; Besides the constraint time/memory/power consumption issue, recent studies [93, 99] indicate that unlike face identity recognition, applying deep CNN to age estimation can not improve the accuracy [141]: While CNNs typically excel at learning hierarchical visual cues, along with the deep feature extraction pipeline CNNs gradually lose localized subtle but critical features e.g., wrinkles and dark skin dots.

How can we effectively leverage feature learning approach to improve the state-of-the-art convolutional feature map aging framework, but still avoid subtle feature loss and unnecessary time/memory/power budget? Our solution is motivated by Convolutional Sparse Coding (CSC)’s recent success in computer vision and its unique advantages: CSC considers the whole image during training instead of patch level, therefore the filter learned by CSC can capture high-order image statistics [142]. Therefore, we employ CSC as the building block to automatically learn a semantically meaningful filter bank, in the hope that the learned filter bank can capture more comprehensive and accurate aging features; Furthermore, CSC runs at the much cheaper time/memory/power budget due to the avoidance of deep layers.
To this end, we propose S-Age, a lightweight aging featuring learning framework. The aging features are learned in an unsupervised manner through CSC. To our best knowledge, this is the first time that CSC is applied to the age estimation framework. The experimental results demonstrate that learned aging features are more discriminative and robust than the reported state-of-the-art aging features, such as bio-inspired features and Gabor filters [86, 83, 105, 91, 96, 98]. More impressively, we find through visualization of the filter, that we can obtain filters that characterize not only low-level but very complicated non-linear aging patterns like eye, nose and wrinkle aging filters (Fig. 5.2). Finally, we successfully control the amount of weights to only 10K.

Our pipeline also have the following contributions: (1) To further improve convolutional map aging framework’s accuracy, we intuitively propose that as a pooling response, Standard Deviation in a pooling window applied in convolutional sparse coding scheme has stronger correlation to age patterns than Max Pooling or Average Pooling (Winkle pattern is a typical example). Therefore, we introduce Standard Deviation Pooling (STD) to the framework to aggregate the convolutional feature maps. Our experimental results verify our hypothesis that STD pooling is more effective than other kinds of pooling algorithms. (2) We apply discriminative manifold learning to the obtained learned features for the purpose of seeking more discriminative low-dimensional representations (reduce the dimension from 50176 to 30) and also improve computational efficiency.

While we target on a constraint mobile phone context, our framework generates impressively accurate result. We evaluate over two standard aging benchmark datasets, i.e., FG-NET and MORPH dataset. By only using our newly learned features we can already outperform the state-of-the-art results. The combination with our STD pooling further improves the performance, which makes our approach excel the state-of-the-art results with a large margin (Table 5.2).

In this work, we focus on exploring a new aging feature. The paper is organized as
An illustration of the proposed framework. Before feeding into the proposed pipeline, images are first aligned based on eyes coordinates. Then a given image goes through the feature extraction pipeline consisting of convolution, Abs, LCN and STD pooling. Next, manifold learning is applied to the vectorized feature maps obtained from previous steps. Manifold learning is employed to transform the high-dimensional feature vector into a low-dimensional feature space, where dimension-reduced features are more discriminative for age estimation. Finally, regression model is used to predict the age label for the input image.

The proposed approach is illustrated in Section 5.2, where we present a detailed description of the proposed scheme which mainly includes the convolutional sparse coding and manifold learning schemes. In Section 5.3, we discuss the dataset and experimental setting. Analysis of experimental results and comparisons with previous methods are also given in Section 5.3. We conclude with a brief summary in Section 5.4.

**5.2 Proposed Approach**

Figure 5.1 illustrates the proposed age estimation framework. Our model takes an entire face image as input. The approach first applies the learned CSC filter bank to the image and generates a set of feature maps. Then the feature maps are processed through a series of nonlinear operations, *i.e.*, absolute value rectification (Abs), local contrast normalization (LCN) and pooling summarization. Next, for efficient computation, the feature vector is projected to a low dimensional space via discriminative feature embedding. Finally, the
5.2.1 Convolutional Sparse Coding

Effectively representing the facial aging pattern is critical for accurate age estimation. However, this type of subtle information is usually very difficult to capture. Previous work on age estimation mainly used hand-crafted features to represent the facial aging characteristics. One representative framework is using bio-inspired features [65, 133, 139, 141, 138], which are a set of predefined Gabor filters with a variety of scales and orientations. However, due to the fact that Gabor filters are essentially edge detectors, their convolution feature maps basically are only sensitive to edges in an image, and therefore they have limited the capability of capturing higher order image statistics. Recently, convolutional sparse coding (CSC) has received increasing research interest in computer vision and machine learning communities and it has achieved very promising results in a variety of problems, such as object recognition [143], pedestrian detection [144] and image segmentation [145]. The reason of CSC’s success is that compared to Gabor filters, the filters learned by CSC are intrinsically shift-invariant and can capture high-order image statistics [142]. As shown in Figure 5.2, the CSC filters learned over MORPH dataset [129] can characterize not only low-level edge information but also very complicated facial patterns, e.g., eye-like filter, nose-like filter, wrinkle-like filters, etc. Therefore, we are motivated to employ CSC [143] as the building block to automatically learn a semantically meaningful filter bank, in the hope of the learned filter bank can capture more comprehensive and more accurate aging features.

Given a training set \( X = \{x_i\}_{1 \leq i \leq N} \), which consists of \( N \) 2D images with dimension \( p \times q \), our goal is to learn a 2D convolutional filter bank \( F = \{f_k\}_{1 \leq k \leq K} \) that can capture the intrinsic patterns within \( X \). Here \( F \) contains \( K \) filters and each filter \( f_k \) is an \( w \times w \) convolutional kernel. We define \( M = \{M^i\}_{1 \leq i \leq N} \) as the set of sparse feature maps such that each subset \( M^i = \{m^i_k\}_{1 \leq k \leq K} \) consists of \( K \) feature maps for reconstructing image \( x_i \).
Figure 5.2: The learned 7 × 7 filters. We highlight a set of impressive filters that are semantically meaningful, e.g., nose (green), eye (red), lip (blue) and wrinkle(purple)-like filters.

where $m^i_k$ has dimension $(p + w - 1) \times (q + w - 1)$. We also define $\|A\|_1$ as the element-wise matrix $\ell_1$-norm, i.e., $\|A\|_1 = \sum_{i,j} |a_{ij}|$, where A represents any 2D matrix and $a_{ij}$ is an element in A at location $(i, j)$. Convolutional sparse coding (CSC) is capable of generating such a filter bank that can characterize higher-order image statistics [143], by decomposing each training image $x_i$ as the sum of a series of sparse feature maps $m^i_k \in M^i$ convolved with kernels $f_k$ from the filter bank $F$. Formally, CSC solves the objective function

$$\min_{F,M} \mathcal{L} = \sum_{i=1}^{N} \left\{ \left\| x_i - \sum_{k=1}^{K} f_k * m^i_k \right\|_F^2 + \alpha \sum_{k=1}^{K} \| m^i_k \|_1 \right\}$$

s.t. $\|f_k\|_2^2 = 1, \forall k = 1, \ldots, K$ \hspace{1cm} (5.1)

where the first term and the second term represent the reconstruction error and the $\ell_1$-norm penalty respectively; $\alpha$ is a small regularization constant; $*$ is the 2D discrete convolution operator; and $f_k, \forall k = 1, \ldots, K$ are restricted to have unit energy to avoid trivial solutions. The construction of $F$ is realized by balancing the reconstruction error and the $\ell_1$-norm penalty.

It is well known that the objective of Eq. (5.1) is not jointly convex with respect to $F$ and $M$ but is convex with respect to one of the variables with the other kept fixed [146]. A common approach for optimizing Eq. (5.1) is solving the two variables alternately, i.e., iteratively performing the two steps that first compute $M$ and then update $F$. Iterative Shrinkage
Thresholding Algorithm (ISTA) is used for computing the sparse feature maps $M$. On updating the filter bank $F$, we use the stochastic gradient descent, \textit{i.e.}, estimating the true gradient based on one training sample at a time [143].

5.2.2 Feature Extraction

As illustrated in Figure 5.1, given an input image, the first step of our framework is to apply the convolutional filter bank learned by CSC to extract intrinsic aging features within the image. Figure 5.2 demonstrates that the learned filter bank contains a rich collection of feature patterns. While producing a certain number of Gabor-like filters, our CSC has a greater capability in representing higher-order image statistics, by generating more complex filters, which in our case are learned specifically for characterizing the underlying aging signatures among human face images. The learned convolutional kernels display significant variations, such as different orientations, frequencies and distinct structures. More complicated patterns are also captured, such as circular shaped filters and corner structure at various angles. This is also in accordance with the findings in works [143, 142].

By taking a closer examination of the CSC filters, we can see that 1) the learned convolutional filter bank has well-managed redundancy, \textit{i.e.}, containing only a small number of low-level edge primitives; 2) the convolutional kernels can capture semantically meaningful key facial regions, \textit{i.e.}, eye-like, nose-like and lip-like filters and can characterize the underlying aging signatures, \textit{i.e.}, wrinkle-like filters, etc.; 3) the filter bank is very compact and has a moderate model complexity (# parameters $\sim O(10k)$), which allows practical deployment on mobile devices with limited power source and computational capability. Therefore, CSC can effectively discover intrinsic aging information. In this work, these filters are learned from all training samples in an unsupervised way. Specifically, on preprocessing, facial images are aligned to the same location based on eyes’ coordinates and then cropped to the same size of $64 \times 64$. The filter size used in our later experiment is set to $7 \times 7$. Without the loss of
generality, we assume that the number of filters is \( K \). To capture sufficient aging signatures from face images, we set filter bank size \( K = 256 \).

Upon learning the filter bank, we extract facial features using the proposed framework illustrated in Figure 5.1, where an input image \( x \) is first aligned and convolved with the learned filters \( f_{1 \leq k \leq K} \). Then the obtained feature maps \( \{ m_k \}_{1 \leq k \leq K} \) are processed through three cascaded layers, namely, element-wise absolute value rectification (Abs), local contrast normalization (LCN), and pooling. The Abs layer computes absolute value element-wise in each feature map \( m_k \). Here we denote \( y^a_k = |m_k| \) as the \( k \)-th feature map after the Abs layer. For the purpose of avoiding the cancelation effect in sequential operations, the LCN layer enhances the stronger feature responses and suppresses weaker ones across feature maps, by performing local subtractive and divisive operations within each feature map. We define the feature maps after LCN to be \( Y^l = \{ y^l_k \}_{1 \leq k \leq K} \).

The pooling layer aims to summarize each feature map \( y^l_k \) by partitioning it into non-overlapping windows and extracting the related response from each of the pooling window. In this work, pooling is conducted in each 6 x 6 window. The most commonly used pooling operation is MP (max pooling) \([112, 143, 147]\). To better characterize the aging characteristics, we employ another nonlinear operation — standard deviation (STD) defined as follows:

\[
y = \sqrt{\frac{1}{N_B \times N_B} \sum_{y_i \in \Omega} (y_i - \overline{y})^2},
\]

where \( \Omega \) is a pooling window of size \( N_B \times N_B \) and \( \overline{y} \) is the mean response in the neighborhood \( \Omega \). We use \( Y^p = \{ y^p_k \}_{1 \leq k \leq K} \) to represent the set of \( K \) feature maps after pooling. Compared to max pooling that preserves only the maximum response within a window, the STD pooling result is computed based on averaging the variance of all responses within the window. Therefore, STD pooling has a better capability in capturing the local variability.
among the responses within each pooling window. In contrast to other classification problems (e.g., face recognition, object categorization), the distinctiveness among different ages is much more subtle. Typically, the visual cues for distinguishing the age differences rely on the local details within facial images, such as wrinkles and dark skin spots (senile plaque). Based on the aforementioned observations, we are motivated to explore the capability of STD pooling, in the hope of extracting more effective aging features. To the best of our knowledge, this is the first successful attempt combining CSC and STD pool and achieving superior results on age estimation. We show in the experiment using STD pool leads to better performance than using max pooling.

Finally, the aging feature for an input image is computed by vectorizing each pooled feature map and then concatenating the obtained vectors. The aging feature vector is represented as \( z = [z_1^T, z_2^T, \ldots, z_K^T]^T \), where \( z_k^T \) denotes the \( k \)-th sub-vector obtained by vectorizing \( k \)-th pooled feature map \( y_k^p \).

### 5.2.3 Discriminative Feature Learning

Upon finishing the previous feature extraction steps, we apply manifold learning to the obtained feature vectors, for the purpose of seeking discriminative low-dimensional representations and improving computational efficiency. [87]. We define \( Z = [z_1, \ldots, z_N] \in \mathbb{R}^{d \times N} \) as the feature matrix \( Z \), containing \( N \) \( d \)-dimensional feature samples. Our goal here is to find a projection matrix \( P \in \mathbb{R}^{d \times l} \) such that the original features \( Z \in \mathbb{R}^{d \times N} \) are transformed to a low-dimensional space \( W = P^T Z \in \mathbb{R}^{l \times N} \) (\( l \geq d \)), in which the new features \( W \) are more discriminative the task being considered. In this work, Orthogonal Locality Preserving Projections (OLPP) [125] is used.

The OLPP aims to find the embedding that preserves manifold structure by measuring the local neighborhood distance. We define three matrices, i.e., symmetric matrix \( S \), diagonal matrix \( D \) and Laplacian matrix \( L = D - S \). If feature vector \( z_i \) is one of the \( k \) nearest
neighbors of $z_j$, the affinity weight $S_{i,j} = \exp(-\|z_i - z_j\|^2/t)$ where $t$ is a positive constant; otherwise, $S_{i,j} = 0$. We let $D_{ii} = \sum_j S_{ij}$. The objective function of OLPP is:

$$\min_P \sum_{i=1}^{N} \sum_{j=1}^{N} \|P^T z_i - P^T z_j\|^2 L_{i,j}$$

s.t. $P^T Z D Z^T P = I$ \hspace{1cm} (5.3)

In age estimation, due to significant face shape variations and large age differences, it is usually very difficult to achieve accurate face alignment. This issue hampers many existing age estimation algorithms that are based on manifold learning [148]. Fortunately, as pointed out in [143, 144, 142], the features learned by the convolutional sparse coding possess a good property of translation invariance, which can greatly mitigate the problem caused by face misalignment. Based on this observation, we believe that the synergy from applying manifold learning to the features learned via CSC can lead to performance improvement in age estimation, because of two advantages: 1) feature extraction becomes insensitive to images misalignment yielding robust feature vectors; 2) better feature samples allows more effective discriminative learning.

5.3 Experiments

5.3.1 Evaluation Criteria

As we know, age estimation can be considered as a regression problem [127, 128]. In this work, we use support vector regression (SVR) to predict the age based on the feature vector obtained from manifold learning. We use two different measuring criteria to evaluate the performance of the proposed framework in age estimation. They are the Mean Absolute Error (MAE) and the Cumulative Score (CS). The MAE is calculated based on the average of the absolute errors between the estimated age and the ground truth (labeled age), which is represented as: $\text{MAE} = \frac{1}{N} \sum_{n=1}^{N} |g_n - g'_n|$, where $g_n$ represents the ground truth label of
Figure 5.3: Overview of age distribution of two datasets (Morph on the left and FG-NET on the right) used in our experiment. X-coordinate represents age label. Y-coordinates represents the number of people within each age category.

Table 5.1: Illustration of the used datasets. \( N_{in} \) indicates the number of individuals. \( Age_r \) represents the age range of the corresponding dataset.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>( N_{in} )</th>
<th>( Age_r )</th>
</tr>
</thead>
<tbody>
<tr>
<td>FG-NET[130]</td>
<td>1002</td>
<td>0-69</td>
</tr>
<tr>
<td>MORPH [129]</td>
<td>5475</td>
<td>16-77</td>
</tr>
</tbody>
</table>

\( nth \) image, \( g'_n \) denotes the estimated age and \( N \) is the total number of testing samples. The Cumulative Score (CS) is calculated as \( CS_g = \frac{C'_g}{C} \times 100\% \), where \( C'_g \) represents the number of probe facial images whose absolute error between the estimated age and the ground truth age is not greater than \( g \) years. \( C \) is the total number of testing samples. We perform experiments with 20 different settings of \( g \).

5.3.2 Datasets and Experimental Settings

To compare with state-of-the-art approaches and evaluate the proposed framework’s performance, two available public benchmark datasets are used, which are the FG-NET [130]
and the MORPH database \cite{129}. Table 5.1 and Figure 5.3 list detailed information about these datasets. FG-NET \cite{130} includes 1002 facial images belonging to 82 different individuals with large variations in expression, illumination and pose. Several samples are listed in Figure 5.4. To our best knowledge, the FG-NET dataset covers the widest age range from 0 to 69 among current popular age datasets. In addition, the images of this dataset have different qualities since some images are taken contemporarily in color while some others are captured a long time ago in gray. Figure 5.4 illustrate some example images from the FG-NET dataset. Although the samples included in the FG-NET dataset are limited, but it has a very wide age range and has become a general benchmark for comparison among different approaches. There are the major reasons why FG-NET is still widely used to evaluate the performance of different age estimation algorithms. Considering the identity influence, there are 82 different folds in the whole experiment. Leave-one person-out (LOPO) setting is applied to the FG-NET dataset, this is the same as \cite{131, 83, 88, 104, 91, 99, 97, 96, 98}. Within each fold, samples belonging to one specific individual are used for testing. The remaining 81 individuals are used for training. The final result is the average result of 82 folds.
MORPH (Album 2) [129] includes more than 50,000 images with different races. Many works have shown that the aging process varies a lot among different races [133, 77, 138]. To alleviate the influence due to an ethnicity, a subset of MORPH is selected for age estimation [88, 104, 99]. This subset only includes Caucasian people as indicated in Figure 5.5. In this chapter, we use the same subset of images and experimental setting as [88, 104, 99]. For the MORPH database, the whole selected database is randomly divided into two different parts. 80% of the data is used for training and 20% is used for testing. There is no overlap between the training and testing sets. The whole experiment is also repeated 30 times as in work [88, 104]. The final result is the average value of results obtained in different experimental folds.

5.3.3 Experimental Results and Analysis

In this section, we evaluate the proposed framework and compare it with several popular models in the literature. Table 5.2 lists the MAE results obtained by the proposed model and other published results on the two benchmark datasets. From the listed results, we can see that by exploiting the CSC features, both of our approaches outperform state-of-the-art
Figure 5.6: (a) Illustration of the facial image distribution of one individual at different ages, on the age manifold for FG-NET dataset. (b) Illustration of facial image distribution of different individuals with different ages on the age manifold for the MORPH dataset. Different color represents different age categories. The number in red denotes the ground truth age of each facial image.
Table 5.2: Comparison of different methods for age estimation on two benchmarks in MAE (years). Proposed scheme indicates STD pooling is applied in the proposed pipeline.

<table>
<thead>
<tr>
<th>Method</th>
<th>Year</th>
<th>FG-NET</th>
<th>MORPH</th>
</tr>
</thead>
<tbody>
<tr>
<td>LARR [83]</td>
<td>2008</td>
<td>5.07</td>
<td>-</td>
</tr>
<tr>
<td>SVR [83]</td>
<td>2008</td>
<td>5.66</td>
<td>5.77</td>
</tr>
<tr>
<td>BIF [65]</td>
<td>2009</td>
<td>4.77</td>
<td>-</td>
</tr>
<tr>
<td>MTWGP [132]</td>
<td>2010</td>
<td>4.83</td>
<td>6.28</td>
</tr>
<tr>
<td>OHRank [88]</td>
<td>2011</td>
<td>4.85</td>
<td>5.69</td>
</tr>
<tr>
<td>PLO [89]</td>
<td>2012</td>
<td>4.82</td>
<td>-</td>
</tr>
<tr>
<td>CPNN [141]</td>
<td>2013</td>
<td>4.76</td>
<td>-</td>
</tr>
<tr>
<td>CA-SVR [104]</td>
<td>2013</td>
<td>4.67</td>
<td>5.88</td>
</tr>
<tr>
<td>Feature-Fusion [91]</td>
<td>2014</td>
<td>4.49</td>
<td>-</td>
</tr>
<tr>
<td>CS-LBFL [98]</td>
<td>2015</td>
<td>4.43</td>
<td>4.52</td>
</tr>
<tr>
<td>C-IsLPP [90]</td>
<td>2013</td>
<td>4.38</td>
<td>-</td>
</tr>
<tr>
<td>Han et al. [96]</td>
<td>2014</td>
<td>4.80</td>
<td>-</td>
</tr>
<tr>
<td>CS-LBMFL [98]</td>
<td>2015</td>
<td>4.36</td>
<td>4.37</td>
</tr>
<tr>
<td>CNN-DLA [99]</td>
<td>2015</td>
<td>4.26</td>
<td>4.77</td>
</tr>
<tr>
<td>CSC+Average Pooling</td>
<td>ours</td>
<td>4.18</td>
<td>3.80</td>
</tr>
<tr>
<td>CSC+Max Pooling</td>
<td>ours</td>
<td>4.10</td>
<td>3.78</td>
</tr>
<tr>
<td>CSC+STD Pooling</td>
<td>ours</td>
<td>4.01</td>
<td>3.66</td>
</tr>
</tbody>
</table>
Figure 5.7: Correlations between age estimates by the proposed approach vs. ground truth ages on (a) Morph and (b) FG-NET.

results over the two datasets by a large margin. Specifically, the combination of our proposed CSC feature and traditional max pooling (baseline approach) yields MAE 4.10 and 3.78, exceeding the best reported results by \(|4.26 - 4.10| = 0.16\) and \(|4.37 - 3.78| = 0.59\) over the FG-NET and the MORPH datasets respectively. Moreover, combining CSC feature with our proposed STD pool further reduces the MAE to 4.01 and 3.66, leading the best reported results by \(|4.26 - 4.01| = 0.25\) and \(|4.37 - 3.66| = 0.71\), respectively, over the two datasets considered. Furthermore, we evaluate the estimation performance using average pooling. Average pooling achieves very similar results as max pooling but still worse than the proposed STD pooling scheme. For completeness, we test the performance using traditional sparse coding with STD pooling, which however only yields 4.50 on FG-NET and 5.10 on MORPH. It is also worthy to note that the proposed framework generates better results compared to Convolutional Neural Network (CNN) based approach [99]. One possible reason is that while CNN is typically good at learning hierarchical visual cues via feature abstraction, one potential drawback of it is gradually losing spatial resolution along the deep
feature extraction pipeline. Aging information, however, are localized subtle features, e.g., wrinkles and dark skin dots. We list the estimation results within different age categories and also compare with one recent work [141], experimental results demonstrate a significant improvement over previous work [141] on six different age categories. We show the correlations between the estimated ages and ground-truth ages in Figure 5.7. The experimental results demonstrate the obtained results are promising.

Compared with previously reported results, the significant improvement confirms the superiority of convolutional sparse coding in modeling aging characteristics over hand-crafted features or the combination of such type of features (HOG, SIFT, Gabor Filters) [91]. Inspection of the features learned by the proposed feature extraction pipeline suggests that our model is more “similar” to the bio-inspired human perception mechanism than traditional bio-inspired features [65, 141]. Like previous works [83, 65, 141, 91, 98, 99], our filters contain Gabor-like low-level edge primitives, as shown in Figure 5.2. More importantly, compared to previous works, they also capture semantically meaningful key facial regions, i.e., eye-like (marked by the red rectangle ), nose-like (marked by the green rectangle ) and lip-like (marked by the blue rectangle) filters as well as the underly aging signatures, i.e., wrinkle-like filters, etc. The most direct reason for this is that our convolutional kernels are directly learned from the data instead of using predefined filters. In addition, the filter bank employed in our experiment contains only $7 \times 7 \times 256 = 12,544$ parameters, which makes the proposed CSC feature detector particularly suitable for practical deployment on computation and power limited hand-held devices, e.g., smart phones, tablets, smart appliances, etc. In this work, we also provide the comparison results using different pooling strategies including STD, MP, AP (average pooling) as shown in Table 5.2, Figure 5.9 and Figure 5.8. We can find that STD obtains a better performance than MP, AP. As far as know, this is the first time that STD nonlinear operation is introduced to the convolutional sparse coding scheme. This result illustrates that standard deviation performs better in capturing aging
Table 5.3: Comparison of MAE in Different Age Ranges on the FG-NET Database

<table>
<thead>
<tr>
<th>Age Range</th>
<th># Images</th>
<th>Proposed</th>
<th>CPNN [141]</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-9</td>
<td>371</td>
<td>2.90</td>
<td>2.30</td>
</tr>
<tr>
<td>10-19</td>
<td>339</td>
<td>2.71</td>
<td>3.83</td>
</tr>
<tr>
<td>20-29</td>
<td>144</td>
<td>4.05</td>
<td>8.01</td>
</tr>
<tr>
<td>30-39</td>
<td>70</td>
<td>6.14</td>
<td>17.91</td>
</tr>
<tr>
<td>40-49</td>
<td>46</td>
<td>11.09</td>
<td>25.26</td>
</tr>
<tr>
<td>50-59</td>
<td>15</td>
<td>14.25</td>
<td>36.40</td>
</tr>
<tr>
<td>60-69</td>
<td>8</td>
<td>26.82</td>
<td>45.63</td>
</tr>
<tr>
<td>Average</td>
<td>1002</td>
<td>4.01</td>
<td>4.76</td>
</tr>
</tbody>
</table>

information.

To further understand the effectiveness of the obtained aging feature and investigate the low-dimensional space found by manifold learning, we also illustrate the distribution of different facial images on the learned 2D manifold based on the learned aging feature. Specifically, Figure 5.6(a) demonstrates that the images of an individual at different ages distribute smoothly along the manifold. In addition, Figure 5.6(b) shows the distribution of facial images of different people at different ages. From the visualization, we can see that 1) the features learned by convolutional sparse coding can effectively extract discriminative aging information; 2) the OLPP manifold learning algorithm can capture very well the intrinsic aging feature even when the feature dimension is reduced to only 2 dimensions.

5.4 Conclusion

In this chapter, we have proposed a new framework for image-based facial age estimation. For the first time, unsupervised learned aging features via convolutional sparse coding is introduced to the age estimation problem. To better capture aging cues, STD pooling is applied to the feature maps generated by convolutional sparse coding and achieves better results than traditional pooling schemes (Max and Average). Manifold learning is
Figure 5.8: Cumulative scores (at error levels from 0 to 20 years) of different schemes on MORPH. STD, AVE, MAX indicate standard deviation, average and max pooling strategy respectively.

used to find the discriminativeness of features in a dimension-reduced space. Our system is evaluated on two benchmark datasets and significantly outperforms the state-of-the-art on both datasets.
Figure 5.9: Cumulative scores (at error levels from 0 to 20 years) of different schemes on FG-NET. STD, AVE, MAX indicate standard deviation, average and max pooling strategy respectively.
Chapter 6

HOW TO MINIMIZE INFLUENCE DUE TO GENDER AND ETHNICITY FOR AGE ESTIMATION

6.1 Introduction

As we discussed in previous chapters, age estimation is a very challenging problem. Many factors restrict the performance of an age estimation system. As analyzed in psychology, aging process includes both intrinsic and extrinsic processes [149]. The aging process also varies among different races and genders [150]. As illustrated in Figure 6.1, we find that even within the same age category, individuals with different genders or races still deviate in facial appearance. To develop a comprehensive age estimation system, these factors have to be considered. In this chapter, we focus on the study of age estimation across different ages and genders.

Compared to works in general age estimation problem, there are relatively few works specifically analyzing the influence of race and gender. Ni et al. studied the problem of cross-database age estimation in [94]; This work used the data collected from internet as training. FG-NET [130] and MORPH [129] were used for testing. Recently, Guo et al. demonstrated that gender and race information are two main factors for age estimation in [133]. They showed that age estimation system usually can get very good performance within the same race and gender group. Compared to these works, our method is different and novel in that it uses correlation learning to jointly mitigate race and gender influences. Furthermore, our study gives a comprehensive analysis of the influence in age estimation due to race and gender variations.
Our major contributions of the proposed framework include: (i) We give a complete and quantitative analysis of age estimation across race and gender. (ii) We propose a new framework to estimate age across different races and genders. This proposed scheme is new for solving this kind of problem. (iii) Our experiments are conducted on a large widely used database (MORPH-II). Excellent performance has been obtained based on the proposed scheme.

In this chapter, we study the influence due to race and gender difference in age estimation. We have investigated the feasibility of applying a machine learning strategy to
**Figure 6.2:** Illustration of the proposed learning scheme. Different colors represent the features extracted from different races (red for African American, and green for Caucasian). The race differences are marked by different shapes, which are circle and triangle. Same size corresponds to the same age label. Our scheme is separating different ages “further” away while keeping same age “closer” regardless of the race (or gender) difference.
diminish these influences. To solve the given problem, we evaluated age estimation performance across two races - Caucasian and African-American. We use MORPH-II data set [129]. We conduct a set of comprehensive experiments to validate the proposed method. Experimental results demonstrate that the proposed estimation scheme can reduce the effect caused by the variation of race and gender. We use the resulting Mean Absolute Error (MAE) to demonstrates the effectiveness of the strategy. This also proves the feasibility of age estimation across different race/gender age groups. In general, this work is a necessary supplement for the current age estimation system. In general, our proposed scheme is a critical component in developing a comprehensive aging estimation system in the future.

The whole chapter is organized as the following: Section 6.2 talks about the proposed scheme for age estimation across races and genders. Then, two major learning schemes, discriminative and correlation learning are discussed. In section 6.4, experimental results are given and discussion are presented in section 6.5.

6.2 Approach

In this section, we give the details of the proposed scheme. First, we apply discriminative subspace learning algorithm to the original aging feature during the training stage. As analyzed in previous studies [87, 133, 105], subspace learning has been proven to significantly reduce the dimensionality redundancy of the original image feature space. We apply this approach to minimize the difference of facial features extracted from the facial images at the same age. Furthermore, with consideration of the race and gender factors, we apply correlation learning in the second step. Our idea is that using correlation learning based scheme can minimize the difference of aging patterns of different races and genders at the same age. At the final step, a regression approach is used to obtain the estimated age for the given probe face image.

In the testing phase, we first extract the feature from the given cropped aligned face
images. Afterwards, we project features into the discriminative and correlation subspace as illustrated in Figure 6.2. Then we use Linear Support Regression as the estimation scheme to obtain the age of a given face image. In the following sections, the algorithms of our scheme are discussed in more details.

6.2.1 Aging Feature Representation

In age estimation problem, how to represent the aging cues efficiently is still an open problem. In [84], facial part ratios and wrinkles presence were calculated to represent the aging feature. Geng et al. built a projection subspace using the sorted time sequence to model the aging feature in [82, 86]. Yan et al. [151] used DCT transform to extract feature in the extracted facial patches. Bio-Inspired Feature (BIF) has been proven to be an efficient aging representation in [65, 102, 139].

In this work, we use BIF feature to represent the facial age characteristics. Bio-Inspired feature is designed to simulate the human visual system into different layers and is widely used in object recognition work. These layers are arranged in a hierarchical structure [85]. In general, there are two adjacent layers included in this model. They are simple (S) and complex (C) cell units. The first layer $S_1$ is computed by the convolution between the given image and Gabor filters. To build the Gabor filter set, four orientations and sixteen scales are employed in [85]. Two adjacent $S_1$ units are used to form 8 new units within each orientation. $C_1$ layer is calculated by applying the maximum pooling within the local spatial area. This procedure simulates the outputs in human early visual system. Recently, a few works are proposed to modify the BIF feature extraction scheme. In [152], two additional layers $S_2$ and $C_2$ are added onto $S_1$ and $C_1$ layers. Most significant changes are taken in $S_2$ layer, instead of convolution, patch matching using the pre-defined category patch is conducted among $C_1$ units to get the feature for $S_2$. The scheme of obtaining $C_2$ from $S_2$ is the same as $C_1$ from $S_1$ units. In [65, 94], only $C_1$ from $S_1$ are used to extract the aging
feature. Their experimental results demonstrate the ineffectiveness of extracted $S_2$ and $C_2$ features based on pre-learned prototypes for age estimation [65]. In this framework, we also employ $C_1$ layer as the representation of aging feature. This is similar as [65, 94]. As we know, although BIF feature has been proven to be a good representation of aging feature, it still cannot represent the age changes among different races and ages. We give the analysis and experiment illustration in the experimental part.

### 6.2.2 Discriminative Aging Mapping

In many applications of pattern recognition and computer vision, such as face recognition and gender classification, dimensionality reduction is usually a necessary step to extract the discriminant and efficient feature representation before classification. In my framework, we use Marginal Fisher Analysis (MFA) as the discriminant learning tool. This approach was original proposed by Xu et al [153]. In general, MFA is used to represent the similarity of intra-class by computing the sum of distances between its sample and its neighbors. In general, the difference among data within inter-class is measured by the sum of distance between margin points and the neighboring points of different classes. The main advantage of using MFA is that it does not only characterize the difference among classes, but also represents the local manifold structure. It is also good at dealing with margin samples [153].

### 6.2.3 Correlation Learning

After obtaining the projecting discriminant subspace using BIF feature via manifold learning scheme, we propose to learn the correlation mapping among different races and genders within the same age category. We plan to reduce the influence due to race and gender variations. Its goal is to make the aging feature representing the same age category obtained from different races and genders “similar”. In this chapter, we evaluate the performance of two different correlation learning algorithms in dealing with this problem: Canonical
Correlation Analysis (CCA) and Partial Least Squares (PLS). Although CCA and PLS have been used as the classification engine in age estimation work [102, 137, 26], as far as we know, no work has applied these two algorithms to learn the correlation of aging features between different races and genders.

6.2.3.1 Canonical Correlation Analysis

Canonical Correlation Analysis (CCA) [154] is a widely used algorithm for multivariate data analysis. CCA is usually used to construct the subspace to maximize the correlation between two sets of related variables. In the high-dimensional space, the correlation between two sets of variables can be modeled as linear relation. This can be represented using linear CCA. Suppose that the corresponding data set is constructed by \((X, Y)\), which represents \(N\) pairs of samples \((x_i, y_i)\). Application of CCA scheme is to learn the projection matrix \(w_x\) and \(w_y\) to maximize the correlation between the projected matrix \(w_x^T x\) and \(w_y^T y\). This can be represented as maximizing the following correlation equation:

\[
v = \frac{w_x^T C_{xy} w_y}{\sqrt{w_x^T C_{xx} w_x w_y^T C_{yy} w_y}}, \tag{6.1}
\]

where \(C_{xx}\) and \(C_{yy}\) indicate the intra-class covariance matrices. \(C_{xy}\) and \(C_{yx}\) indicate the inter-class covariance matrices. This equation can be simplified as maximizing \(v^2\), which can be represented as

\[
\text{corr}(x', y') = \frac{\text{cov}(x', y')}{\text{var}(x') \text{var}(y')}, \tag{6.2}
\]

where \(x'\) indicates \(w_x^T x\), and \(y'\) indicates \(w_y^T y\). \(\text{cov}(x', y')\) is the covariance value between \(x'\) and \(y'\). In this chapter, we evaluate the effectiveness of CCA in the work of correlation learning. Meanwhile, we also compare the performance of CCA with PLS in the following section. In our work, \(x\) and \(y\) indicate the aging feature of the same age across different races.
6.2.3.2 Partial Least Squares

Partial Least Squares (PLS) [155] is a regression model. The major function of PLS is to project the given variables (input) and the predicted response (output) to a new low dimensional latent linear space. The covariance between the input latent scores and the response is maximized in this new low dimension subspace. It is accomplished by computing the linear mapping from input variables’ latent score to response.

Let us assume that the two corresponding datasets are $X \in \mathbb{R}^M$ and $Y \in \mathbb{R}^N$, where $M$ and $N$ indicates the dimension of $X, Y$ space separately. To compute the correlation between $X$ and $Y$, PLS can be computed as follows:

\[
X = VP^T + E
\]
\[
Y = UQ^T + F,
\]

(6.3)

$V$ and $U$ are $l \times d$ matrices of the $d$ extracted latent vectors or PLS scores. $l$ represents the number of samples. Matrix $P \in \mathbb{R}^{M \times d}$ and matrix $Q \in \mathbb{R}^{N \times d}$ are loading matrices. $E \in \mathbb{R}^{l \times M}$ and $F \in \mathbb{R}^{l \times N}$ are residual matrices. The relation between matrices $V$ and $U$ can be modeled as

\[
U = VD + H,
\]

(6.4)

where $D$ indicates $d \times d$ diagonal matrix relating with the latent scores of $X$ and $Y$. $H \in \mathbb{R}^{l \times d}$ is the residual matrix. PLS tries to calculate the normalized basis vectors $w$ and $c$, then maximize the covariance between PLS score vectors $t$ and $u$ (corresponding row vector in $V$
and \( U \). This can be calculated as

\[
\max([\text{cov}(v, u)]^2) = \max([\text{cov}(Xw, Yc)]^2),
\]

(6.5)

where \( w \) and \( c \) are the normalized vectors which satisfy \( ||v|| = ||c|| = 1 \) and \( \text{cov}(v, u) = \frac{v^Tu}{n} \) is used to measure the covariance between \( t \) and \( u \). With iteration of this calculation process, we can obtain basis which is used for projecting variable \( X \) and \( Y \) to a new subspace. After combining Eqn. 6.3 and Eqn. 6.4, \( Y \) can be represented as

\[
Y = UQ^T + F = (VD + H)Q^T + H \\
= VDQ^T + (HQ^T + F).
\]

(6.6)

We can simplify the equation as

\[
Y = VP^T + O,
\]

(6.7)

where \( P = QD^T \) is the regression matrix \( n \times d \) from \( X \) to latent space, and \( O = HQ^T + F \) is the residual matrix. Based on the illustration shown above, we find PLS is more “complete” than CCA. As illustrated in [156], one shortcoming of CCA is that it does not capture the character of these unseen testing points, which means it cannot capture the difference between these training data under certain situations.

Compared to CCA, PLS is widely used to compute the regression function from the feature space to a label space for performing classification [102]. Recently, PLS is also applied to solve cross-modality face recognition problem [156]. In their work, two correlated variables \( X \) and \( Y \) are used to represent the same individual’s features extracted from two different modalities. Motivated by their idea, we use PLS to learn the correlation between individuals’ aging features obtained from different races/genders at the same age.
6.2.4 Classification (Aging Function)

After feeding the discriminative aging feature into the correlation learning scheme to get the correlation aging feature, the classification engine is employed to map the projection pattern to the specific age label. In our work, Support Vector Regressor (SVR) [106] is used to estimate the age for the input given face image.

6.3 Dataset

It is known that FG-NET [130] and MORPH [129] are two widely used datasets for age estimation. Because of the distribution of images across different ethnicities in FG-NET is highly skewed and limited in size (1002), we choose MORPH-II (more than 55,000 faces images) as the evaluation database in our framework for evaluating the performance of the proposed scheme. The images included in MORPH-II database have a large ages span, from 16 to 67 years old. The attributes information is also provided in this database. These information includes age, gender and race. The labeled age information is especially important for the quantitative analysis of the proposed scheme.

As illustrated in Table 6.3, we can find that the percentage of Caucasian and African-American images in the whole dataset is more than 95% of the entire database. This is illustrated in Table 6.1. Also, the number of Caucasian and African-American subjects is not balanced. The number of African-American subjects is more than 4 times larger than the Caucasian. With consideration of unbalanced face images distribution among different races in MORPH-II, we use all Caucasian individuals. Then we choose a random selection of African-American faces from the whole database according to the number of Caucasian faces as shown in table 6.3. This resulting dataset is more balanced compared to the original dataset with regards to racial distribution.
### Table 6.1: The distribution of face images with different races and genders in Morph II database

<table>
<thead>
<tr>
<th>Race</th>
<th>Male</th>
<th>Female</th>
<th>Total Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caucasian</td>
<td>7999</td>
<td>2601</td>
<td>10600</td>
</tr>
<tr>
<td>African-American</td>
<td>36803</td>
<td>5757</td>
<td>42560</td>
</tr>
<tr>
<td>Hispanic</td>
<td>1651</td>
<td>100</td>
<td>1751</td>
</tr>
<tr>
<td>Asia</td>
<td>146</td>
<td>13</td>
<td>159</td>
</tr>
<tr>
<td>India</td>
<td>43</td>
<td>14</td>
<td>57</td>
</tr>
<tr>
<td>Other</td>
<td>3</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>All</td>
<td>46645</td>
<td>8487</td>
<td>55132</td>
</tr>
</tbody>
</table>

### 6.4 Experiment

In this section, we evaluate the performance of the proposed scheme in dealing with the influence caused by race and gender. We use mean absolute error (MAE) and cumulative score (CS) to evaluate the performance of age estimation using different algorithms. We use $g_k$ and $g'_k$ represent the ground truth age and estimated age separately. $K$ is the total number of test images. MAE is calculated as

$$MAE = \frac{1}{K} \sum_{k=1}^{K} |g_k - g'_k|.$$  \hspace{1cm} (6.8)

The cumulative score (CS) is represented as follows:

$$CS_r = \frac{K'_r}{K} \times 100\%$$  \hspace{1cm} (6.9)

In this equation, $K'_r$ is the number of probe face images whose absolute error between the estimated age and the ground truth age is not greater than $r$ years old.
6.4.1 Experiment Setting

For the purpose of studying the influence of age estimation scheme across races and genders, we arrange our experimental setting into four cases: no cross race/gender, cross race, cross gender and cross race and gender. In no cross race/gender, the samples used for training and testing are collected from the same race and same gender. We divide the whole dataset into four groups: Caucasian male (WM), Caucasian female (WF), African-American male (BM) and African-American female (BF). Then we also equally divide the data of each specific group into two parts: $P_1$ and $P_2$. In our experiments, $P_1$ and $P_2$ are interchanged as training and testing sets. There is no overlaps between the training and testing data in all our experiments. Then we have four different subgroups (WM,WF,BM,BF) and four different types of experiments (no cross race/gender, cross race, cross gender, cross race and gender). In general, there will be 16 different experiments needed to be conducted in this work. Two-fold cross-validation is applied in our experiment as follows:

**No Cross race/gender experiment:** For example, when we evaluate the age estimation result of Caucasian male images, there are two subgroups $WM_1$ and $WM_2$. Initially, $WM_1$ is used for training and $WM_2$ is used for testing, $WM_1 \rightarrow WM_2$. Then, in the second fold, the groups for testing and training are swapped: $WM_2 \rightarrow WM_1$. The final MAE is the average value of these two experiments.

**Cross experiment:** The datasets used in the experiment are from different age face categories. For instance, in one experiment, we use Caucasian male group ($WM$) for training, and African-American male group $BM$ for testing. There are four kinds of experiments including $WM_1 \rightarrow BM_1$, $WM_1 \rightarrow BM_2$, $WM_2 \rightarrow BM_1$ and $WM_2 \rightarrow BM_2$. The final result is calculated by the average of these four experiments’ result.
Table 6.2: Experimental Results: MAE(in yrs.). B+M+C represent BIF feature + MFA + CCA. B+M+P represents BIF+MFA+PLS. The age groups include African-American Female(BF), African-American Male(BM), Caucasian Female (WF). MAE Incr. indicates the percentage of advance using B+M+P compared to the result only based on BIF.

<table>
<thead>
<tr>
<th>Train</th>
<th>Test</th>
<th>BIF</th>
<th>BIF+MFA</th>
<th>B+M+C</th>
<th>B+M+P</th>
<th>MAE Incr.(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BF</td>
<td>BF</td>
<td>5.41</td>
<td>5.12</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>BM</td>
<td>8.63</td>
<td>8.35</td>
<td>8.31</td>
<td>4.93</td>
<td>42.87</td>
</tr>
<tr>
<td></td>
<td>WM</td>
<td>8.83</td>
<td>7.34</td>
<td>6.32</td>
<td>5.80</td>
<td>28.43</td>
</tr>
<tr>
<td></td>
<td>WF</td>
<td>7.87</td>
<td>6.93</td>
<td>6.61</td>
<td>5.50</td>
<td>16.01</td>
</tr>
<tr>
<td>WM</td>
<td>WF</td>
<td>4.54</td>
<td>4.17</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>WM</td>
<td>8.47</td>
<td>7.10</td>
<td>6.12</td>
<td>5.25</td>
<td>38.02</td>
</tr>
<tr>
<td></td>
<td>BM</td>
<td>8.96</td>
<td>8.82</td>
<td>7.58</td>
<td>5.60</td>
<td>42.87</td>
</tr>
<tr>
<td></td>
<td>BF</td>
<td>7.98</td>
<td>7.25</td>
<td>6.79</td>
<td>5.84</td>
<td>42.87</td>
</tr>
<tr>
<td>WM</td>
<td>BM</td>
<td>4.65</td>
<td>4.30</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>BF</td>
<td>8.29</td>
<td>8.29</td>
<td>8.04</td>
<td>5.87</td>
<td>29.19</td>
</tr>
<tr>
<td></td>
<td>WM</td>
<td>6.88</td>
<td>6.33</td>
<td>6.64</td>
<td>5.20</td>
<td>24.42</td>
</tr>
<tr>
<td></td>
<td>WF</td>
<td>9.87</td>
<td>7.93</td>
<td>6.55</td>
<td>5.62</td>
<td>43.06</td>
</tr>
<tr>
<td>BM</td>
<td>WM</td>
<td>3.90</td>
<td>3.82</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>WF</td>
<td>6.81</td>
<td>6.58</td>
<td>6.52</td>
<td>5.38</td>
<td>21.00</td>
</tr>
<tr>
<td></td>
<td>BM</td>
<td>7.23</td>
<td>6.98</td>
<td>6.39</td>
<td>5.42</td>
<td>25.03</td>
</tr>
<tr>
<td></td>
<td>BF</td>
<td>7.71</td>
<td>7.20</td>
<td>6.93</td>
<td>6.03</td>
<td>21.79</td>
</tr>
</tbody>
</table>

Table 6.3: The distribution of images used in our experiment

<table>
<thead>
<tr>
<th></th>
<th>Male</th>
<th>Female</th>
<th>Total Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caucasian</td>
<td>7999</td>
<td>2601</td>
<td>10600</td>
</tr>
<tr>
<td>African-American</td>
<td>7999</td>
<td>2601</td>
<td>10600</td>
</tr>
</tbody>
</table>
6.4.2 Experimental Results Analysis

To evaluate the performance of our proposed scheme, we need a baseline experiment. Our proposed scheme is to minimize the influence due to race and gender variations for age estimation. So no-cross experiment is used as our baseline to measure the effectiveness of the advocated algorithm. Based on the obtained results, we find that the BIF feature is a good representation for human age characteristics. It performs very well in age estimation within the same age group. However, MAE significantly increases when the experiments are conducted across different races and genders. For instance, using WF as training and WM as testing, the obtained MAE is 8.47 compared with 4.54 gotten within the Caucasian Female group. These results indicate that the aging process significantly varies among races and genders. This result is also in accordance with the analysis illustrated in [150].

In addition, we find that race variation has more influence on the age estimation result under most cases compared to gender factor. Discriminant learning can help improve the age estimation performance. But when the experiment is evaluated between different age groups, the improvement is not “significant” any more. It also reflects that discriminant learning does not have the enough “capability” to learn the correlation among different features. However, after applying correlation learning, the performance significantly improves. Even for some experiments across different age groups, the MAE is even smaller than the results obtained between the same age group, such as from BF to BM. After integrating the discriminative mapping and correlation learning, MAE is 4.93. This is smaller than MAE result estimated from BF to BF (5.41).

As listed in Table III and Figure 6.3, we find that the proposed scheme can significantly reduce the influences from race and gender variations. Experimental results demonstrate that PLS performs better than CCA in dealing with this problem.
6.5 Conclusion

In this chapter, we have investigated the problem of age estimation across races and genders. We have conducted extensive experiments across all different age groups on a large database. The results demonstrate that race and gender information plays an important role in age estimation, especially race variations. We propose a new learning scheme to mitigate the influences brought by race and gender variation. We also present a strategy based on discriminative and correlation learning that can diminish the influence due to the variations of races and gender. Two different correlation learning schemes (PLS and CCA) are also analyzed and compared. Our obtained results demonstrate that the proposed approach is effective in minimizing the influence brought by gender and race in age estimation. This work is a critical study towards developing a comprehensive age estimation system with race and gender diversity.
Figure 6.3: The cumulative scores of age estimation based on different learning schemes. No cross means the training and testing sets are within the same race/gender age group. Except no cross, all other experiments are conducted between different age groups for training and testing. Each curve is the average of the experimental results of four ethnicity/gender age groups (WF, WM, BF, BM).
Chapter 7

SUMMARY

Analyzing social relations through facial images has attracted lots of interests in the computer vision and image processing communities. In this thesis, I propose a new framework by leveraging the facial geometry and appearance information for kinship verification. The appearance model is based on the idea of seeking familial traits from two facial images. Instead of calculating facial similarities using two global faces directly, I aim to find familial traits from facial patches based on Gaussian Mixture model. Then the similarity is calculated from these corresponding facial patches where familial traits are located. One interesting discovery is that a geometry model is very helpful for kinship verification. The proposed geometry model uses landmarks extracted from facial images as the basis to measure the similarity of facial geometries. With the integration of the geometry model, I have obtained an additional 5% improvement compared to the result obtained by the appearance model only. Experimental results demonstrate that both facial appearance and shape information convey useful cues for kinship verification.

Not limited to measure the kinship relation between pairwise facial images, my works study the problem of automatically detecting family photos from the group photo gallery. I utilize three different cues to represent the characteristics of group photo. The first model is based on the standing positions of people. It is designed to capture the standing geometry of people. The proposed geometry model can capture the overall geometry of face locations at the scene level although purely utilizes the relative positions of people in the image. This makes the process of geometry feature extraction very efficient and leads to satisfactory
performance gain in family photo categorization problem. To capture the people’s facial similarities, I have proposed two different kinds of appearance features. The major difference between two different schemes is the way to extract facial appearance descriptor. The first scheme is based on the traditional appearance features-SIFT. The second scheme uses deep neural network to extract feature descriptor. To measure the facial similarities of group people, I propose a new mid-level feature Degree of Group Similarity Feature (DOGF) where age information is incorporated in the proposed feature extraction model. This is used to minimize the influence of age gap for facial similarity measurement. Experimental results illustrate the importance of age information in capturing the group photo characteristics. In addition to geometry and appearance information, I also use semantic information to separate two categories. Finally, a simple fusion scheme is applied to combine all the three cues together. Combining all the above factors gives significant performance improvements. I have conducted experiments on a dataset consisting of thousands of group photos downloaded from Flickr. While any feature representation could be valid for the task, one important observation is that different feature representations actually carry complimentary information from different aspects.

Due to the importance of age estimation in the computer vision community and its essential role in group relation analysis, I propose a new supervised framework for extracting aging features. The proposed framework is based on the Convolutional Neural Network (CNN) using supervised learning strategy. Taking advantage of CNN’s powerful capability in feature representation from images, in this thesis, I apply CNN to extract aging feature from static facial images. Different from previous models using CNN, I use feature maps obtained from different layers instead of using the feature obtained at the top layer. The extracted feature has demonstrated its strong discriminative power for age estimation problem.

To reduce the complexity of the age estimation model, I have proposed a new framework for image-based facial age estimation. This work is motivated by the success of
previous convolutional aging feature (e.g., bio-inspired features (BIF)) which obtains very promising age estimation results. One drawback of BIF is that the manually-crafted filters cannot easily capture the complicated facial aging patterns. In this work, I adopt similar convolutional map framework but using a novel feature learning approach based on Convolutional Sparse Coding (CSC). The framework adopts the Standard Deviation Pooling (STD) to summarize the aging feature, which yields better aging feature correlation compared to the traditional average pooling and max pooling. Finally, the extracted features are fed into a discriminative manifold learning model to obtain more discriminative low-dimensional representations and further improve the computational efficiency. The evaluation on two standard benchmark datasets demonstrates that our method outperforms the state-of-the-art methods.

The performance of age estimation is usually influenced by many factors, such as different races and genders. In this thesis, I investigate how the age estimation system can be influenced by two important factors: race and gender. The discovery is that: race variation has more influence on the age estimation result compared to the gender difference. In the beginning, I test the performance of discriminative mapping. Although it can help improve the age estimation performance, the improvement is not evident when the experiments are performed across different age groups. I apply correlation learning to map the features into another subspace which can maximize the correlation of feature samples at the same age between different domains. The proposed scheme has demonstrated its efficiency in mitigating the influences brought by race and gender variations.

To summarize, the main theme of my thesis is leveraging the different image feature learning techniques and machine learning models to help predict and understand the social relations between individuals. Towards this goal, I have explored the methods for pairwise kinship verification and group people relation analysis (family); to leverage important age information, I have proposed two different aging estimation frameworks using both supervised and unsupervised aging feature learning frameworks; to alleviate the influence of race
and gender variations for age estimation, I have developed a framework to incorporate the discriminant feature mapping and correlation learning. I believe these techniques will be very helpful in understanding the social connections from images and therefore bring more insights to understand the social behaviors in social media and improve the user engagement and ads recommendations for a much smarter social media ecosystem.
BIBLIOGRAPHY


[56] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: A unified embedding for face recognition and clustering, CVPR.


C. M. Bishop, et al., Neural networks for pattern recognition.


L. I. Smith, A tutorial on principal components analysis, Cornell University, USA 51 (2002) 52.


